PROCEEDINGS

THE MILITARY TESTING ASSOCIATION

MTA

21st Annual Conference of the

MILITARY TESTING ASSOCIATION

Coordinated by the

NAVY PERSONNEL RESEARCH & DEVELOPMENT CENTER

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

nprdc

SAN DIEGO, CALIFORNIA

SAN DIEGO, CALIFORNIA
15 - 19 OCTOBER 1979

12

PROCEEDINGS

21st Annual Conference
of the
MILITARY TESTING ASSOCIATION

Coordinated By:

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
San Diego, California

and

BAHIA MOTOR HOTEL
Mission Beach, California

15-19 October 1979

A

FOREWORD

The papers presented at the Twenty First Annual Conference of the Military Testing Association came from the business, educational, and military communities, both foreign and domestic. The papers reflect the opinions of their authors only, and are not to be construed as the official policy of any institution, government, or branch of armed service.

TABLE OF CONTENTS

21st Annual Conference of the
MILITARY TESTING ASSOCIATION

Officers, Chairpersons, and Committee

MTA President
CAPTAIN DONALD F. PARKER

MTA Chairperson
MARTIN F. WISKOFF, PhD

MTA Secretary
PAUL P. FOLEY

Chairpersons:	Committee:
NORMAN ABRAHAMS, PhD	Program
TED YELLEN	Financial
RANDY STOTLAND, PhD	Audiovisual
JOHN ELLIS, PhD	Social
BARBARA STANEK	Registration
BART KUHN	Public Relations Officer

21st Annual Conference
MILITARY TESTING ASSOCIATION

15 October - 19 October

MONDAY, Oct. 15

0800 - 1700 Mezzanine (Bldg 600) - Registration

1600 - 1800 Pacific Room (614) - Steering Committee Meeting

	MISSION BAY ROOM	DELMAR ROOM	MISSION ROOM	BAY ROOM
TUESDAY, Oct. 16				
0900 - 1000	Conference called to order			
0900 - 0915	Greetings by Capt. Donald F. Parker			
0915 - 1000	Keynote Address by RADM James R. Hogg			
1000 - 1030	-break-			
1030 - 1130	Air Traffic Controllers I			
1130 - 1140	Announcements			
1140 - 1300	LUNCH			
1300 - 1430		Air Traffic Controllers II	Selection & Testing	Training Issues
1430 - 1500		-break-	-break-	-break-
1500 - 1630		Air Traffic Controllers II*	Adaptive Testing	Training & Testing
1700 - 1900		Social Hour		
WEDNESDAY, Oct. 17				
0800 - 0930		Contingency Approach to Management	Occupational Analysis I	Soft Skill Analysis
0930 - 1000		-break-	-break-	-break-
1000 - 1130		Management & Survey Research	Occupational Analysis II	Nutrition & Health
		LUNCH	LUNCH	LUNCH
1300 - 1430		Assessment Center Executive Devel.	Occupational Analysis III	Training Effectiveness
1430 - 1500		-break-	-break-	break-
1500 - 1630		Officer Selection, Training, & Eval.	Occupational Analysis IV	Training Task Analysis I
THURSDAY, Oct. 18				
0800 - 0930		Personnel Classification & Assignment	Evaluating Training Performance	General Topics
0930 - 1000		-break-	-break-	-break-
1000 - 1130		Recruiting/Assess. of Enl. Personnel	Training Task Analysis II	Selection & Validation
1130 - 1300		LUNCH	LUNCH	LUNCH
1300 - 1430		Issues in Selection & Validation	Officer Task Analysis I	Issues in Criterion-Referenced Testing
1430 - 1500		-break-	-break-	-break-
1500 - 1630		Women's R&D **	Officer Task Analysis II	Performance Testing
1730 - 2200		Social Hour & Luau		
FRIDAY, Oct. 19				
0800 - 1000	Computer-Aided Career Information			

*Session closes at 1615

**Session closes at 1600

PAPERS IN THEIR ORDER OF PRESENTATION

PAPER SESSION: Occupational Analysis (I)

PAPER SESSION: Occupational Analysis (II)

SYMPOSIUM: Soft Skill Analysis
Soft Skill Analysis in the Army
Chairperson: LTC Bradford L. Walton
Training Developments Institute

SYMPOSIUM: Nutritional and Health Issues among Military Personnel
The Impact of Nutritional Status on Health Risk Factors and Job Performance
Chairperson: Owen Maller, PhD
U.S. Army Research and Development Command

PAPER SESSION: Issues in Selection and Validation
Chairperson: John F. McAreavy, PhD
Chairperson: John F. McAreavy, PhD
U.S. Army Management Engineeiing Training Activity

PAPER SESSION: Women R & D Issues

DISCUSSION SESSION: Officer Task Analysis (I)

SYMPOSIUM: Officer Task Analysis (II)
How Thin Can Officer Job Analysis Baloney Be Sliced?
Chairperson: Francis F. Medland
U.S. Army Research Institute

MTA KEYNOTE ADDRESS

Rear Admiral J. R. Hogg

Director, Military Personnel and
Training Division

16 October 1979

KEYNOTE ADDRESS TO THE MILITARY TESTING ASSOCIATION

16 OCTOBER 1979

INTRODUCTION

I am very pleased to have this opportunity to address you at this 21st Military Testing Association meeting. This forum has served over the years as a valuable means for professionals, such as yourselves, to share research experiences and evaluate their usefulness. As a member of the user community, I want to emphasize the importance of your contributions and to assure you that we will rely on your continued services as we make our decisions in the future. In that context, I would like to discuss four areas which may be useful to you in developing your products. These can be categorized as: policy oriented; establishing institutional memories; designing innovative programs; and, interchange of findings and ideas.

By being policy oriented, I mean that both managers and researchers must avoid thinking in terms of the present. Managers have a tendency to focus on the matter of immediate impact without considering long term trends and consequences. You, as researchers, can help us in this regard by taking a "full spectrum" approach to the projects under your cognizance. This means that the issues at hand must be addressed, while also looking downstream to their long term implications. For example, the decision to lower recruiting entry standards may net us an immediate gain in meeting end strength while causing a long term negative impact on the enlisted force quality mix. While this may be an obvious oversimplification, the press of circumstances in particular cases could easily result in a decision to embark upon a course of action without full appreciation of its ramifications. I ask that you think along these lines; by doing so, you will be of increased assistance to management decision-makers.

I would further recommend that you make it standard practice to remain with a project throughout its implementation phase. The gains for the manager are the value of advice and counsel of the program developer when it is time to implement. This also helps lend impetus to a program, which otherwise may be lost in the hassle of daily crises after receipt of the initial report. Many good ideas have come to all stop, due to lack of follow-through in the implementation process. Related to this is the benefit derived from using good display and selling techniques when it comes to presenting proposed programs to decision-makers. I am not advocating that each of you become a salesman. Rather, my point is that, in our environment, it is not enough for you to be excellent researchers. It is extremely valuable to managers to have the full breadth of a new program outlined in a manner that is logical, which highlights pros and cons, and reaches realistic conclusions and practical recommendations upon which action can be taken. In other words, you must catch our attention and make us listen until program implementation is complete!

My next point, establishing institutionalized memories, is close to the interests of all historians and others with 20/20 hindsight. We waste a lot of time and repeat a lot of mistakes by not remembering and learning from past experiences. In the case of research projects, this translates into maintaining a complete file of background data on all projects, cross-indexed for easy reference. No doubt, this is already a standard part of your procedures. However, managers are not always aware of the presence of such useful information. Therefore, I urge you to surface such data, so that we do not spend a lot of time reinventing the wheel. An example of how such historical information may be brought to bear on current problems relates to the selection of Naval Academy Midshipmen. Data accumulated from the mid-60's has been extremely beneficial to us in recent years. We have continually expanded on the early data base and applied it with great success in meeting our increasing requirements for more technically oriented officers of the future. In summary, the maintenance of ready reference sources with details of assumptions applied to specific cases, maxmizes the efficient utilization of your research talents. In addition, it may lead to quicker decisions which can be applied in a timely manner in pursuing our objectives.

Now to the designing of innovative programs. I believe there is much to be gained from taking some calculated risks -- by you, in pushing the state of the art, and by managers, in their willingness to embark on pilot tests of new programs. An example of this is our recently established Job Oriented Basic Skills, or JOBS, program which I had the pleasure of inaugurating in its pilot stage at NTC San Diego in July. The increasing technical demands of our weapons systems drives the need for a large number of personnel who are eligible for vocational training in high skill areas. However, the market's availability of qualified applicants is diminishing. To correct this situation we were able, with significant research help, to develop the JOBS program. JOBS is applied to sailors in boot camp who are initially ineligible to attend our Class "A" Schools. It augments their basic skills and prepares them to complete follow-on technical training with a high rate of success. Program success to date leads us to believe we can expand the program to all of our RTCs in October 1980. Had we not been willing to be innovative, we would never have realized the benefits of such an approach.

Another strong program has evolved from a successful recruit training film, which was an R&D effort in the Marine Corps. The desire was to better prepare new recruits for their experiences in recruit training. Our 80 minute film is shown to new recruits shortly after they report for basic training.

It is designed to acquaint them with training, and relieve their uncertainty and apprehension. We believe this will reduce attrition in recruit training and provide us with sailors who are ready to hit the decks in their first shipboard assignments. Later, I will show you an excerpt from the film.

In the interchange of findings and ideas, as I have already noted, we can learn from each other and from other branches of the service. We are long past the point where we can afford parochialism. There should be no unions in the sharing of information and ideas, for we are all in the same sobering business of defending our country. The Air Force has been a leader in designing innovative enlisted classification procedures. Through mutual cooperation and the efforts of NPRDC, Navy soon will go on-line with an updated computerized enlisted classification system called CLASP, or "Classification and Assignment within PRIDE." We believe this will provide a better match of our personnel with job requirements, to the benefit of both the individual and the Navy. The Army's Military Aptitude Profile, or MAP program, has been instrumental in Navy's development of an applicant screening program based on biographical information. This will enhance our ability to assess the potential of our prospective recruits. Navy's recently developed recruiter selection system is being evaluated by Army in conjunction with a recruiter assessment center they are developing. Army feels they can easily adapt our program to their needs. In summary, we can and need to learn from each other.

Now, here is the recruit training film excerpt I mentioned earlier (10 minute film clip).

I would like to talk now, based primarily on my Navy experiences, about two areas in which I feel we have special mutual interest and concern. The first is continued development of ASVAB. CAT has great promise as a long-term solution, but it is still on the horizon; therefore, refinement of ASVAB is our best near and mid-term solution for basic entrance tasking. In that regard, I was disappointed to learn -- upon returning to Washington duty last April -- that implementation of ASVAB 8, 9, and 10 has been delayed from late-1979 until May 1980. When my service counterparts, MEPCOM, and I, as members of the DOD ASVAB Steering Committee, carefully reviewed the situation, we realized that the realistic implementation date -- due to several developmental factors -- is now late September 1980. I can assure you that OSD, the Services, and MEPCOM are putting a lot of emphasis on making that date a reality, without further slippage.

The second area is realted to the nation's decreasing number of qualified military applicants as we move into the 1980's. The predictions made in the 70's are now becoming a reality. Demographic reviews for 1980 and 81 show a real decline in numbers and we all know the forecast beyond that. Navy did not meet recruiting goal in 1979, falling short some 5,000 recruits, primarily in the general detail area. What we cannot afford, given the declining number of available recruits, is the attrition of our young men and women, once recruited, but prior to their completion of obligated service. We simply must improve in our ability to screen military candidates in terms of motivation and attitude, in addition to mental aptitude. In fact, I'll go a step further and say, based on my observation of sailors in the fleet and the significant challenges they face, that we need to be able to measure such things as the frustration level and degree of flexibility of military candidates. This is not possible today, and it may not be possible in the future, but it is the challenge that I respectfully place before you; because, these and related personal behavior traits are most certainly the difference between success and failure in the Navy and, I expect, throughout the military and in fast moving private organizations as well. You can see that I am going beyond RBQ and MAP in this discussion. There is no other choice for, in my opinion, given our current ability to recruit, and the market availability for the next 20 years, we must develop this more sophisiticated screening capability as a matter of necessity to ensure our ability to maintain the required numbers and quality of personnel in the military.

In closing, I realize I have done a lot of preaching to the choir. However, I want you to know that I am personally convinced of the great importance of concerted and shared R&D efforts. Your contributions are essential if we are to triumph over the manpower, personnel and training problems confronting us, and we must do that for the sake of the military and our country. There is much to be done in tailoring our recruiting programs to the current and potential markets. Those applicants we accept must be equipped with a good chance of career success. Then, too, we must devote our attention to our mid-grade officer and enlisted forces to ensure their continued professional growth, personal satisfaction and contributions.

I appreciate the opportunity to be here today and to share some of my thoughts with you. You can be assured of my continued interest in the R&D program and of my personal participation in working together with you in addressing future issues. Thank you.

SELECTION AND EVALUATION OF AIR TRAFFIC CONTROLLERS

Historic Overview of Research and Development for Air Traffic Controller Selection

by Leland Brokaw, Ph.D.

Introduction

This paper will cover selected research studies accomplished between 1952 and 1972; dealing generally with the development of predictors for use in the selection of Air Traffic Control (ATC) personnel. Although I have indicated citations for specific quotations, I wish to acknowledge the debt I owe to the writers of a number of reports which have provided source material. Mahlon V. Taylor in his comprehensive summary of American Institute for Research contract efforts in the early 50's provides a suitable landmark for the initiation of this report. In 1956, when the Civil Aeronautics Administration Air Traffic Control school visited the Air Force for further assistance in development of a selection screen, I was priviledged to initiate an experiment which treated school criteria and job performance 1 year after graduation as possible measures for evaluation of background, experience, and test predictors. In 1961, David Trites, at the Civil Aeromedical Institute, picked up the data base and further validated the experimental measures after controllers had been 5 years on the job. The late Bart B. Cobb, in the late 60's, further investigated age and experience variables, and revalidated the tests which had been adopted by FAA for selection use before 1964. He further evaluated the test battery for Marine Corps and Navy air traffic control personnel. The late W. Dean Chiles, in the early 70's, expanded the research base to include a computer based multiple task performance predictor which demonstrated significant validity.

Discussion

The improved selection of air traffic control personnel has been a matter of interest for at least the last 38 years. Increased air traffic, technological change, and increasing need for specific job relevance have each contributed to the need for additional studies during this time.

Mahlon V. Taylor in his 1952 report of a research effort by the American Institute for Research (AIR) cites an unpublished study accomplished by Dewey Anderson in 1941 in which 28 tests, including 17 Thurstone Primary Mental Ability tests were given to most controllers in service. Against performance ratings the four highest validities were developed by a measure of perceptual speed and accuracy entitled "Three Higher;" a space relations test called "flags," and tests of reasoning and integration entitled "Letter Series," and "Pedigrees." These factor areas were included in the battery assembled by AIR in 1951 in their accomplishment of a contract written by the Civil Aeronautics Administration calling for the development and validation of a test battery designed to screen applicants for jobs in Air Traffic Control.

As an historic footnote, Dr. John C. Flanagan supervised the AIR contract effort, with the advice of Dr. Elmer D. West and Miss Marion Shaycroft; Paul M. Fitts, Jr., served as research adviser and rendered useful criticism throughout the project. L. Dewey Anderson and Bryce Hartman were consultants at early stages of the project.

In addition to the measures designed to pick up the factorial areas identified in the earlier work, additional tests were prepared to address the content areas identified in a job analysis, as well as tests designed to serve as job samples. In addition to the job analysis, AIR personnel solicited critical incidents from supervisors in both air route control centers and control towers. The critical activities and accompanying aptitude components identified by AIR are given in Table 1. It must be noted that a stated requirement for paper-and-pencil measure for use by the Civil Service Commission made it impossible to measure auditory perception, writing rapidly and legibly, copying behind, speaking intelligibly at optimal speed, verbal fluency, and the long term aspect of memory.

Eighteen tests were developed to measure the aptitude components as they might occur in real job situations. After preliminary tryout and refinement the test battery was administered to 211 persons, 90 in air traffic control centers, 75 in control towers, and 46 from communication stations. Applied research in an operating environment often has difficulty with samples for study. This work is no exception. The 90 persons from centers were tested in five locations, the 75 from towers in 12 locations, the 46 from communication stations in seven places. As a result there are many small samples making collection of criterion data critical. Criterion losses made it impossible to validate the communication samples. A process of validating within each subsample, and deriving average validity correlations through Fisher's z-transformation provided analytic samples of useful size.

A number of kinds of data were collected during the criterion development. Biographical data and existing official evaluations were obtained for personnel of the facilities at which tests were given. In addition, supervisors were requested to record incidents of critical performance on the Airways Operations Performance Record which had been developed for the job analysis, and to provide overall performance records of personnel taking the tests. Quoting from Taylor (1952) "In general all the measures were useful as criteria, although there was considerable variability among facilities in this respect, and it was felt advisable to select and combine, for each of the seven facilities, the measures (aside from supervisor's ratings) proving most satisfactory for the given facility. The correlations between supervisors' ratings and these composite criteria averaged .81, indicating that both....were measuring the same thing."

As indicated above, when the tests were administered in the field practical difficulties made themselves felt. Both personnel to be tested and testing time were in short supply, so that sample sizes adequate for determination of statistical significance of the validities were obtained for only 14 of the tests. Table 2 presents the derived validities for 16 of the tests, but tests numbered 12 and 13 were not considered when the recommended battery of nine was selected.

In summarizing the results of the study, Taylor (1952) indicated that, although the battery was suitable for use with candidates who had prior experience in flying, or in some controller function, he felt that use of the tests with naive candidates would require extensive revision of the directions by experienced test development personnel, and that additional experimental testing would be required.

Although the American Institute for Research delivered a recommended battery of tests for the selection of trainees for the air traffic field, it was never implemented. The next event in the flow of research was a visit by CAA representatives to the Personnel Laboratory, Air Force Personnel and Training Research Center.

Representatives of the Air Traffic Control Branch, Federal Airways Standardization Division, Civil Aeronautics Center, Oklahoma City, visited the Air Force Personnel Laboratory in February 1956 to inquire into methods of selecting control tower and air traffic control personnel. Their visit was prompted by a desire to improve selection procedures for trainees in the CAA Air Traffic Control School (ATC).

At that time their selection was based upon previous on-the-job experience and a physical examination. Need for increasing numbers of trainees led to the requirement for a method of selection from a naive population.

A joint Air Force/Civil Aeronautics Administration study was undertaken, and a battery of 20 tests was administered to entering trainees in the ATC school. The battery was quite heterogeneous, selected from a number of sources to cover the variance believed to be relevant to success in the training. Commercial tests, Air Force tests, and the more effective tests devised by the American Institute for Research were included. Initial validation was accomplished against available school criteria, including an average lecture grade, individual instructor ratings of student proficiency, and a composite instructor rating, based upon discussion between the three or four instructors who had dealt with each class. There were about 20 students in each class: of 197 students receiving the experimental battery, training criterion data were accumulated for 130. The tests are identified, and their validities for the three training criteria are presented in Table 3.

This study was intended to provide a test battery to replace a previous system involving experience variables. In that context, the contribution of experience to school success was of interest. Background and experience variables included in the study were age, education, marital status, and previous air traffic experience. Experience was studied in these five categories:

A. Trainees reporting any air traffic control experience versus those with no experience.
B. Experience in airport traffic control versus no such experience.
C. Experience in ground control approach versus no such experience.

D. Trainees holding Senior CAA ratings versus trainees without such ratings.

E. Trainees with CAA certification versus those without such certification.

The validities of the background and experience variables are presented in Table 4.

Multiple correlations were run to evaluate the contribution of various measures to the prediction of school success. Two criteria were chosen—the average lecture grade and the composite instructor rating. These data appear in Table 5. It should be noted that only the general variable of CAA certification status was sufficiently unique to appear in the multiples. Its zero order validity was of the same order as the test variables, and its contribution was significant.

As reported by Brokaw in 1959 technical note, in May 1957 the trainees who had received the experimental test battery were followed onto the job, and various criteria of their performance collected. Essentially, the same supervisory ratings were used as were collected from the instructors during the training phase of the study. The CAA also collected data on the time spent on the job by each controller before he was recommended for promotion from the trainee level to the helper level. The correlation between the instructor rating during training with the supervisory rating on the job was .59; the average academic lecture grade correlated with the supervisory rating .33. In a sample of 133 controllers these values were both significant at the one-percent level. The period before recommendation to promotion tended to be highly peaked around the third and fourth month on the job. This lack of variance produced a correlation of .18 with the supervisory rating, rendering that criterion value of little use.

The higher validities for experimental tests were found in the mathematical reasoning, computational, and numerical tests. The AIR Air Traffic Problems test showed acceptable validity for both academic criteria and supervisory ratings.

The abstract reasoning and perceptual tests showed useful levels of predictive efficiency for the supervisory ratings, but the verbal and clerical speed and accuracy measures were not significantly related to the job criterion. The background and experience variables tended to be marginally predictive of job performance.

The data base developed by Brokaw in the 1956-57 time period, provided predictor variables for another follow-up by Trites in 1961. The Trites study addressed the extent to which the data would be predictive of current job performance, retention in air traffic control work, incidents of unsatisfactory air traffic control work, and medical history information over the 5 years period.

Quoting from Trites (1961), "Regional offices of the Federal Aviation Agency were able to supply current FAA facility addresses, or other information on all but 10 of the original 197 subjects....Of the remaining 187 subjects, 16 had failed the training course and left the FAA, two were deceased, replies were not received for two, and three were with the FAA but no other information was available. This left 149 subjects (including four training course failures still with the FAA) for whom relatively complete criterion data were obtained."

Criterion data collected included (1) average supervisor rating, (2) active vs. inactive controller, (3) with the FAA vs. not with the FAA, (4) mean hours of sick leave, (5) no symptoms vs. symptoms, and (6) no disciplinary action vs. disciplinary action.

Trites accomplished a number of multiple regression analyses against various combinations of predictors for the listed criteria. He summarized the findings as follows (1961), "....Evaluations of psychological test and biographical data collected when controllers went through training indicated that: (1) Psychological tests can make a useful contribution to screening applicants for air traffic control work; (2) Instructors in the air traffic control school make exceptionally valid predictions of job performance evaluations some years later; (3) Older trainees tended to receive poorer job performance ratings some years later than did their younger classmates; (4) Medical history information of the kind collected in this study is not predictable by the psychological tests which were used."

In the period between 1961 and 1967 American aviation expanded at an accelerated pace. According to Cobb (1967) pilot certification more than doubled between 1962 and 1967, and aircraft production in 1967 was 40 percent greater than in 1965. Growing numbers of aircraft with higher speed capabilities resulted in increasingly heavy work loads for both airport traffic control and air route traffic control.

The increasing demands placed upon controller personnel through the higher volume of higher speed traffic was the basis for growing concern on the part of FAA management regarding the extent to which controller performance and reactions might be associative with aging and length of experience in active control work. Research efforts to answer these questions were initiated in 1965 by the Civil Aeromedical Institute (CAMI) in Oklahoma City.

One aspect of this study was accomplished by Cobb (1967) who accomplished a survey of several hundred journeymen radar control specialists at four Air Route Traffic Control Centers to determine the extent to which job performance might be associative with chronological age and length of experience in control work. For each of several experimentally derived ratings of job performance a statistically significant and negative relationship was found with age. Mean group ratings for controllers over 40 years of age were significantly lower than those of younger groups. Length of experience, when considered independently of age, was found to be of negligible importance and no statistically significant interaction of age and experience were discovered.

Representatives of the air traffic control school at the Glynco Naval Air Station visited the Civil Aeromedical Institute in 1965 to explore methods of improving their selection of Marine Corps and Naval Air Traffic Control trainees. After some discussion, they suggested the use of the operational Civil Service Commission and traffic controller selection battery. Inasmuch as policy considerations made such use of the Civil Service Commission tests impossible, the Civil Aeromedical Institute entered into a joint study with the Navy representatives to assess the validity of commercial tests functionally replicating the Civil Service tests. Such data would be useful to the FAA because they had little opportunity to collect data on sizable samples of naive subjects.

As reported by Cobb (1968) the Civil Service Commission battery consisted of seven tests: Spatial Patterns, Computations, Abstract Reasoning, Letter Sequence, Oral Directions, and Air Traffic Problems. In addition to the current service selection composites, the Civil Aeromedical Institute suggested seven commercial tests for experimental validation. These include the Differential Aptitude Tests of Space Relations, Numerical Ability, and Abstract Reasoning. The others were from the California Test Bureau's Test of Mental Maturity and included Analogies, Inference, Numerical Quantity Coins, and Arithmetic.

The validities of the commercial tests for FAA and military samples, the Civil Service tests for the FAA sample, and the military screening score for the military samples appear in Table 7.

The results indicated that a composite score of four of the commercial tests could be used effectively to predict performance grade and pass-fail status for the Glynco training course. The DAT Space Relations, Numerical Ability, and Abstract Reasoning and CTMM Inference yielded a composite validity of .45 for course grade, and .39 for pass-fail.

The Marine trainees were found to have been selected from relatively higher military screening and classification test scores than were the Navy trainees.

Recommendations coming from Cobb's 1968 study, included the establishment of comparable military screening scores for Navy and Marine samples, with the four test commercial composite reserved for secondary screening, if it became necessary.

Cobb repeated the validation of the seven commercial tests within FAA, Army, Air Force, Marine Corps, and Navy samples in 1971. Results were essentially as in the 1968 study, with the same four test commercial composite being somewhat more valid than the military selection measures used for each sample. In every case, raising the cutoff score on the military screen would have permitted screening at a higher level of effectiveness than would have been possible if the commercial tests had been inserted as a secondary screen.

In 1971 Cobb, Lay, and Bourdet investigated the relationship between chronological age and aptitude test measures within advanced level air traffic control trainees. The study examined the interrelationships of age, aptitude measures and training performance for 710 men who entered basic air traffic control training at the FAA Academy during November 1968 through March 1970. They ranged in age from 21 to 52 years, but less than 12 percent were over 40. Age correlated negatively with 21 of 22 aptitude measures and with training course grades. On most tests, performance means for subjects over age 34 were significantly lower than those obtained for the younger trainees and their attrition rates from training were three times that of their younger classmates. Only one of the 22 measures failed to correlate positively with the training grades. The results indicate that greater effectiveness in screening such applicants could be attained if eligibility standards were modified to include consideration of both age and aptitude.

Previously discussed measures, while valid for prediction of school success and training performance, do not measure a particular kind of ability to be said to be a defining characteristic of a good controller—the ability to perform several different tasks simultaneously. In 1972 Chiles, Jennings, West and Abernathy reported an attempt to assess individual skill at concurrent, time shared performance of a variety of tasks.

They employed the Multiple Task Performance Battery (MTPB), a computer based device intended to measure monitoring, information processing, mental arithmetic, visual discrimination, and inter-individual interaction in the execution of procedures.

Two hundred and 29 air traffic control trainees were tested on Multiple Task Performance Battery. The criterion of trainee potential was based on ratings from FAA Academy instructors in courses being attended by the trainees. Five studies were conducted. The first being in the nature of a pilot study for checking out procedures. The second study (N=60) yielded a validity coefficient of .54. The third study (N=31) yielded a coefficient of .53. The fourth study (N=30) found no predictive power for the MTPB. The fifth study (N=89) produced a coefficient of .24 for one method of computing the performance index and .46 for a second method. For each study the coefficient is based on 1 hour of testing with about 50 minutes of preceding instruction and practice. It is concluded that the MTPB approach to selection offers promise as a screening device for Air Traffic Control Specialist applicants, but further research is required to establish this as a fact and to determine its utility in terms of cost effectiveness.

Summary

This report has covered a number of studies of variables related to success in training and on the job for air traffic control personnel. Throughout the material, it is apparent that spatial visualization, abstract reasoning, and quantitative skills are relevant to the topical area. Credit must be given to those measures which replicate job samples, the American Institute for Research Air Traffic problems tests appeared in the first reported study, and composed a key portion of the Civil Service Commission screen for this specialty as established in 1964.

References

Brokaw, L. D., Selection measures for air traffic control training, Personal Laboratory, Air Force Personnel and Training Research Center, Lackland Air Force Base, Texas. (PL-TM-57-14, July 1957)

Brokaw, L. D., School and job validation of selection measures for air traffic control training, Personnel Laboratory, Wright Air Development Center, Lackland Air Force Base, Texas. (WADC-TN-59-39, April 1959)

Chiles, W. D., Jennings, A. E., West, G., Abernathy, W. T., Multiple task performance as a predictor of the potential of air traffic controller trainees, Federal Aviation Administration, Civil Aeromedical Institute, Oklahoma City, Oklahoma. (FAA-AM-72-5, January 1972)

Cobb, B. B., The relationships between chronological age, length of experience, and job performance ratings of air route traffic control specialists, Federal Aviation Administration, Civil Aeromedical Institute, Oklahoma City, Oklahoma. (FAA-AM-67-1, June 1967)

Cobb, B. B., A comparative study of air traffic trainee aptitude test measures involving Navy, Marine Corps, and FAA controllers, Federal Aviation Administration, Civil Aeromedical Institute, Oklahoma City, Oklahoma. (FAA-AM-68-14, September 1968)

Cobb, B. B., Air traffic aptitude test measures of military and FAA controller trainees, Federal Aviation Administration, Civil Aeromedical Institute, Oklahoma City, Oklahoma. (FAA-AM-71-40)

Cobb, B. B., Lay, C. D., Bourdet, N. M., The relationship between chronological age and aptitude test measures of advanced level air traffic control trainees, Federal Aviation Administration, Civil Aeromedical Institute, Oklahoma City, Oklahoma. (FAA-AM-71-36, July 1971)

Taylor, M. V., Jr., The development and validation of a series of aptitude tests for the selection of personnel for positions in the field of air traffic control, American Institute for Research, Pittsburgh, Pennsylvania, 1952.

Trites, D. K., Problems in air traffic management: I. Longitudinal prediction of effectiveness of air traffic controller, Federal Aviation Agency, Civil Aeromedical Research Institute, Oklahoma City, Oklahoma. (CARI Report 61-1, 1961)

Table 1

Critical activities and Incidents and Aptitude Components for Air Route Traffic Control (R), Airport Traffic Control (P), and Aircraft Communicator (air-ground) (C)[a]

Critical Activity and Aptitude Components	Percent of critical incidents[b]		
	R	P	C
1. Receiving oral messages: auditory perception verbal comprehension	8	13	22
2. Recording oral messages; writing rapidly and legibly, copying behind, encoding, memory for interrupted tasks	9	9	23
3. Recording self-originated data; visual perception, carefulness	9	6	9
4. Displaying flight data; memory for interrupted tasks, visual perception, carefulness	11	8	1
5. Requesting information:	8	11	12
6. Coordinating (Clearances, etc.): memory for interrupted tasks, carefulness, integration, assimilating symbolized data, comprehending unseen movements	21	11	3
7. Devising clearances: carefulness, integration, short term memory, judgment, visual perception of time-distance relationships, assimilating symbolized data, comprehending unseen movement	11	8	0
8. Devising taxiing, takeoff, and landing instructions: carefulness, integration, short term memory, judgment, visual perception of time-distance relationships, assimilating symbolized data, comprehending unseen movement	2	13	0
9. Issuing oral communications: carefulness, verbal fluency, speaking intelligibly at optimal speed	10	10	13
10. Evaluating priority of communications: judgment, integrations, short term memory	7	11	17

a. Taken from Taylor, 1952

b. These are percents of all critical incidents reported for a given job during the test validation (799 for Air Route Traffic Control, 1193 for Airport Traffic Control, 150 for Communications)

Air Route Traffic Control and Airport Traffic Control Average Test Validities with Composite Criteria and With Supervisors' Ratings

Table 2

Test	Composite Criteria		Supervisors' Ratings	
	r	N[b]	r	N
1. Locating Data I	.06	71	.09	83
2. Locating Data II	.06	71	.08	71
3. Air Traffic Math I	.05	71	.06	83
4. Air Traffic Math II	.04	54	.19	54
*5. Memory Flight Information	.20	53	.24	53
*6. Air Traffic Problems I	.49	52	.51	64
*7. Air Traffic Problems II	.53	52	.43	64
*8. Flight Location	.18	62	.32	74
*10. Coding Flight Data I	.29	68	.29	80
12. Taxiing Aircraft	.21	19	.28	19
13. Control Judgment	.39	37	.36	37
*14. Memory for Aircraft Position	.33	76	.22	88
15. Three Dimensional Visualization	.22	27	.08	41
*16. Circling Aircraft	.38	44	.28	56
*17. Aircraft Position	.28	44	.36	56
*18. Flight Paths	.17	44	.37	56

a. Data extracted from Taylor (1952)

b. The rs are as reliable as if computed from a single sample with the given N.

* Tests included in the recommended battery.

Table 3

Validities of Experimental Tests for Three Air Traffic School Criteria[a]

N = 130

Content Area	Test	Validity[b] 1[c]	2	3
Computational and Arithmetic Reasoning				
	Dial and Table Reading (Air Force)	.43	.38	.38
	California Capacity Questionnaire(6)	.17	.22	.24
	Number Series (Mental Maturity)	.21	.18	.23
	Numerical Quantity (Mental Maturity)	.28	.21	.24
	Arithmetic, Personnel Sel and Class	.33	.23	.32
	Numerical Ability, DAT	.32	.28	.31
	Air Traffic Problem I, AIR	.30	.31	.37
	Arithmetic Reasoning (Air Force)	.38	.27	.26
Perceptual and Abstract Reasoning				
	Calif. Capacity Questionnaire (5)	.33	.29	.28
	Abstract Reasoning, DAT	.18	.15	.20
	Space Relations, DAT	.21	.15	.20
	Aerial Landmarks, Air Force	.16	.27	.27
	Spatial Orientation, Air Force	.13	.22	.18
	Instrument Comprehension, Air Force	.26	.23	.23
Verbal Tests				
	Calif. Capacity Questionnaire (7)	.18	.18	.19
	Reading, Personnel Sel and Class	.28	.16	.16
	Language Usage, Sentences, DAT	.22	.18	.15
	Verbal Test, Air Force	.21	.15	.13
Perceptual Speed and Accuracy				
	Code Translation, Calif Test Bureau	.27	.24	.27
	Counting, Calif Test Bureau	.29	.24	.19

a. Taken from Brokaw (1957)

b. Correlations of .17 are significant at the 5% level; .23 at the 1% level.

c. Criterion 1 is average lecture grade, 2 is independent instructor rating of overall performance, 3 is an instructor composite rating based upon the joint judgment of three or four instructors teaching each class.

Table 4

Validities of Selected Background and Experience Factors for Three Air Traffic Control School Criteria[a]

N = 130

Variable	Mean	SD	Validity[b] 1[c]	2	3
Age	26.21	4.20	-.04	-.24	-.24
Education	12.59	1.22	.16	-.11	-.07
Marital Status[d]	.68	.46	.16	.05	.08
Previous Flying Experience	.17	.37	.09	-.10	-.11
Airport Traffic Control	.65	.48	.22	.24	.24
Ground Control Approach	.41	.49	-.06	.15	.14
Any Air Traffic Experience	.78	.41	.04	.23	.24
Senior CAA Rating	.91	.40	.26	.26	.24
CAA Certification in Any Status	.41	.49	.28	.39	.38

a. Taken from Brokaw (1957)

b. .17 significant at the 5% level, .23 significant at the 1% level.

c. Criterion 1 is average lecture grade, 2 is independent instructors' rating, 3 is a composite rating based upon the joint agreement of three or four instructors teaching each class.

d. Marital Status and subsequent entries are dichotomized and computed as point-biserial correlations.

Table 5

Most Efficient Combinations of 2-5 Variables for Prediction of Two Criteria[a]

N=130

Variable	Val	Beta Weights[b]							
		2-var		3-var		4-var		5-var	
Average Lecture Grade									
Arithmetic Reasoning	.43	.38	.38	.33	.30	.36	.31	.28	.30
CAA Certification Status	.32	.23		.22		.26		.23	
Air Traffic Problem I	.30		.21	.20	.22	.27	.29	.29	.26
Symbol Reasoning	.34				.21		.25	.22	.26
Locating Data AIR	-.02					-.22	-.22	-.25	-.24
Code Translation	.24								.15
Multiple Correlation[c]		.49	.48	.52	.51	.56	.55	.59	.57
Composite Instructor Rating									
Air Traffic Problems I	.36	.33	.31	.29	.31	.24	.27	.24	.26
CAA Certification Status	.37	.33		.30		.32		.29	
Arithmetic Reasoning	.32		.28	.24	.22	.20	.20	.17	.21
Symbolic Reasoning	.28				.20		.20	.15	.19
Code Translation	.27					.17	.13	.16	.12
Family Relations	.26								.14
Multiple Correlation		.49	.45	.54	.49	.56	.51	.58	.53

a. Taken from Brokaw (1957)

b. In each pair of columns the first selection is from all variables, the second is from test variables only.

c. All significant at the 1% level.

Table 6

Multiple Validity of Selected Tests for Training and On-the-job Criteria[a]

Test	Instructor Ratings (N=130 Students)		Supervisor Ratings (N=133 controllers)	
	Beta Wt	Validity	Beta Wt	Validity
Air Traffic Problems	.27	.36	.20	.25
Arithmetic Reasoning	.20	.32	.11	.23
Symbolic Reasoning and Perceptual Speed	.20	.28	.16	.22
Code Translation	.13	.27	.05	.14
Multiple Correlation	.51[b]		.34[b]	

Table 7

Validities of Selected Commercial Tests And Civil Service Commission Tests For Military and Civilian Samples[a]

	FAA Sample		Military I		Military II	
	Course Grade	P-F	Course Grade	P-F	Course Grade	P-F
	N_t	N_p-N_f	N_t	N_p-N_f	N_t	N_p-N_f
	r	r_{pb}	r	r_{pb}	r	r_{pb}
Commercial Tests	N=211	193-17	N=635	581-64	N=297	259-35
DAT Space Relations	.14	.20	.31	.18	.25	.12
DAT Numerical Ability	.33	.19	.38	.22	.40	.27
DAT Abstract Reasoning	.40	.38	.29	.14	.36	.23
CTMM Analogies	.12*	.09*	.15	.11	.08*	.04*
CTMM Inference	.27	.25	.24	.21	.13	.12
CTMM N. Q. Coins	.31	.25	.27	.19	.13	.09*
CTMM Arithmetic	.38	.16	.33	.19	.31	.24
Civil Service Commission Tests	N=183	143-40				
Spatial Patterns	.37	.27				
Computations	.28	.16				
Abstract Reasoning	.28	.18				
Letter Sequence	.55	.45				
Oral Directions	.23	.23				
Air Traffic Problems	.41	.29				
Military Scores			N=422	379-51	N=197	171-26
GCT + ARI for Naval trainees			.42	.28	.47	.27
1/3 (AR+VE+PA) for Marine trainees			.18	.01*	.27	.18

a. Extracted from Cobb (1966)

*. Not significant beyond the 5% level

Basic Characteristics of the Air Traffic Controller

by John T. Dailey, Ph.D.

Introduction

In designing a man-machine system for maintaining separation of aircraft in the future, it is very important to examine the characteristics of the basic human raw material that will be fabricated (trained) into the human components of the air traffic control man-machine system. This raw material is not the average human being. The population of air traffic control trainees is a unique group without parallel in other professional or occupational groups. In order to obtain a comprehensive profile of the basic skills and aptitude of two representative groups of air traffic control specialist trainees, a group of 52 entering trainees at the Air Traffic Control Academy were tested on November 19, 1978, with the various Dailey Vocational Tests and another group of 290 air traffic controller trainees were tested in 1970. These tests are designed to measure the potential of your people for a wide range of occupations. They measure a number of the most important skills, found to be associated with preference for training, occupational choice, and performance. The sub-tests measure knowledge of electricity, electronics, mechanics, physical sciences, arithmetic reasoning, elementary algebra, vocabulary and spatial visualization. An electrical composite score is composed of electricity, electronics, mechanics, and physical sciences. A mechanical composite score is composed of mechanics and arithmetic reasoning. A scholastic composite score is composed of arithmetic reasoning, elementary algebra and vocabulary. These tests are modeled after the tests used by the Air Force and Navy for many years for the selection and classification of technical trainees.

The test results were translated into percentile rank by grade 12 male norms. A percentile rank of 71 would mean that the average person in the group was better than 71 percent of the normative group, which happens to be a nationally representative sample of males in the 12th grade. Conversely, a percentile rank of 11 would mean that the average person in that group would be better than only 11 percent of the group in a normative group. By definition, the average score for 12th grade males would be at the 50th percentile in this situation.

It was found that in 1970, air traffic controller students varied tremendously in their relative average level of skill or aptitude in these various areas. They range from the 86th percentile in electronics and mechanics down to the 14th percentile in elementary algebra.

Air traffic controller students in 1970 were exceptionally high in mechanics and also extremely high in electricity and electronics. On the other hand, relative to this high level, they were at a far lower level in such things as physical sciences, arithmetic reasoning and elementary algebra. They were quite high in vocabulary and of course very high in overall score. They were also very high relatively in spatial visualization. It is interesting to note they were the highest group of all in spatial visualization and also tied for highest with airframe repair students in mechanics. This is an extremely interesting finding. They may be extremely high in spatial visualization just because the air traffic controller aptitude tests stress measures of spatial visualization. However, the air traffic controller aptitude tests have no mechanical tests whatever, so this exceptionally high degree of mechanical aptitude must come by the indirect screening of the types of people who want to be air traffic controllers and the kinds of prior experience they have had. Nearly all air traffic controller students in 1970 had considerable prior experience with some aspects of the utilization of complex equipment in the communications, air control or similar fields. This apparently is what generated their exceptionally high level of mechanical comprehension.

They were also extremely high in vocabulary, although the aptitude test did not include measures of vocabulary as such. Apparently, a lot of indirect screening factors tended to generate a group that was very high in vocabulary. The group, predominantly, was a group with little prior college training, and many of the students did not complete high school and have acquired a high school diploma by examination. The low points on the profile in physical sciences, arithmetic reasoning, and algebra are typical for groups without college training.

In Figure 1, one can see the estimated average performances on the total Technical and Scholastic tests of a large number of groups. The scores on the T&S test are estimated from the score on the Information Test of Project Talent. In 1960, Project Talent tested 500,000 high school students and began a series of follow-up studies. In Figure 1 are shown students in various college or noncollege groups 1 year after graduation from high school. There is enormous variation in the average information total level of these groups. For example, the very highest individual scored beyond the 99th percentile. The three high colleges scored at about the 99th or 98th percentile, and highly selective colleges such as the Ivy League scored at about the 89th percentile. The highest three out of about a 1000 high schools average almost as highly as the entering freshman class of the Ivy League colleges. Major state universities were at about the 72nd percentile, with all college students being about 63rd percentile. Junior college students were at about the 53rd percentile, whereas grade 12 students were at about the 42nd percentile. The three lowest high schools average at the first percentile. This means that the senior class at three lowest schools average at the first percentile in terms of grade 12 national norms. The three lowest colleges averaged at about the fourth or fifth percentile. Of primary interest are the noncollege groups shown on the right of the figure. These range from waiter at about the fourth percentile to electronic technician at about the seventieth percentile. As can be seen, the average air traffic control specialist student averaged at about the level of the electronic technician and was only slightly below the level of major state universities. They were above the average for all college students. These are students entering the academy and no attrition had occurred at the time of testing. One year later, undoubtedly, they would score considerably higher since there will be an appreciable amount of attrition in school and on the job and past experience has shown that this attrition is pretty highly related to aptitude level. It can be seen then that the air traffic controller group is an exceptionally high level group aptitude-wise and is at a higher level than the average of all college students. They are considerably higher than the average senior college student.

Figure 2 shows the aptitude profile of the air traffic controller groups as compared to norms for the male 12th graders. Also shown on the chart are results for electronics students in specialty schools, airframe repair students in specialty schools, secretarial students in specialty schools, and a group of broadcasting students in specialty schools. It can be seen that the air traffic controller group in 1970 was a unique group. They had the very high technological aptitude of the electronics and airframe repair students, but also the high verbal level of the broadcasting students.

Experimental Testing of Journeymen Air Traffic Controllers

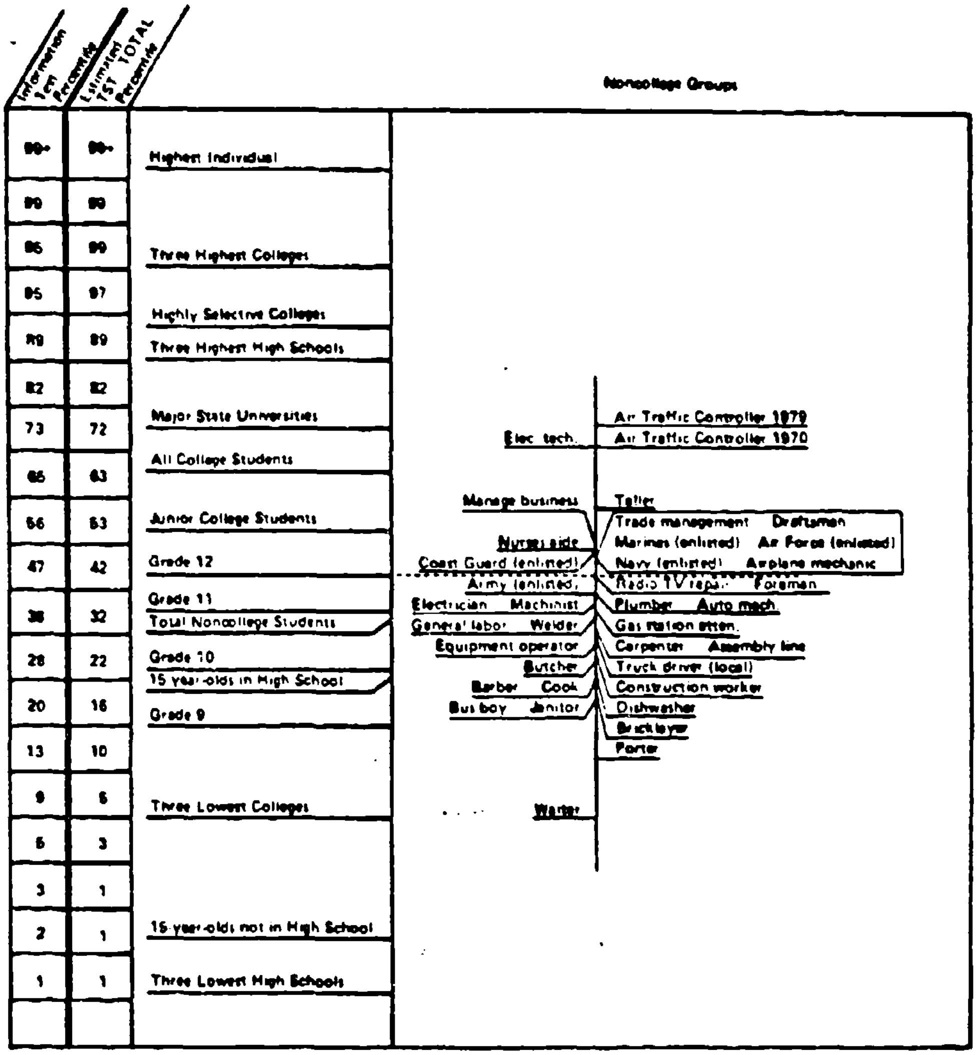
Data are now available on a sample of 140 journeymen air traffic controllers tested at the GS-12 and 13 levels. These data are shown in Table 1 enclosed. This table also shows distributions of aptitude composite scores of those entering the Academy as trainees in 1961 through 1963 and the proportion at each aptitude level that survived to remain on the job 6 years later. It can be seen that at any level some of the trainees will survive training and on the job attrition and still be remaining several years later, presumably doing an adequate job. Those who succeed in training on the job at the lower levels of test aptitude do so because they have additional skills, motivation and other characteristics not measured by the tests. At the very highest level of aptitude, all of the students were still on the job 6 years later. The proportion doing so becomes smaller as score levels become smaller.

Relative to the distribution of aptitudes in the general public, a composite score of 190 is extremely high and only a very small proportion of the general public could do as well as that. However, it can be seen that 80 percent of the air traffic controllers now on the job scored at a level of 190 or above and the median score for incumbents was 226. The current cutting score of 210 is very high. Despite the extremely high level of this standard, approximately two-thirds of the present incumbents are above it. This indicates that air traffic controllers as a group, although they were not specifically screened on the Civil Service Aptitude Tests, are an extremely high level group on these aptitudes. They are as high level aptitude-wise as most professional groups that normally require college graduation for entry. This distribution of test scores by incumbent journeymen as compared to the general population tends to indicate that the tests are appropriate for use as a prescreening standard for air traffic control trainees, and that a relatively high standard for this test seems to be reasonable.

In July-August 1979, the tests were administered to 202 entering ATC students (75 Terminal, 99 Enroute, 28 FAA), the results are shown in Figures 1 and 2. Significant changes are found since 1970. The group in 1979 was considerably lower in electricity, electronics and mechanics. They were considerably higher in science and algebra. They were much higher in spatial visualization and arithmetic and were somewhat higher in vocabulary and in total score. The reason for the changes are unclear, but probably at least in part reflect the policy change in 1973 where the maximum age of 31 was instituted and where GS-9's were no longer recruited with no aptitude test requirement. The 1979 group is definitely better in educational development. This may prove to be a valuable asset to them as they acquire supervisory and administrative responsibilities later in their careers.

After nearly 10 years of experimentation and development of two new types of job related ATC Selection Tests, fully developed multiple operational forms of the Multiplex Controller Aptitude Test and the ATC Occupational Knowledge Test are now available and have demonstrated a high degree of accuracy in predicting Pass-Fail in ATC Academy training and the attainment of journeymen status as a radar controller. Since June 20, 1978, two forms of the MCAT and the OKT have been administered to each entering student on the first day of ATC Academy training. Figures 3, 4, and 5, show the percent of students passing ATC training at each aptitude level (sum of scores on two MCAT forms and OKT). In these figures, Enroute and Terminal students are combined. The composite score on the left is in two columns to reflect the use of two different forms of the Occupational Knowledge Test (Form 101B had 100 items, Form 101C had only 60 items). The composite score reflects nine levels of ability. The number of students is given at each level. At each level is a bar showing the percent of students at that level who pass ATC training at the ATC Academy. The relationship is very sharp. At the upper level all students passed and at the lower level no students passed. The relationship is even sharper for 150 program students and women than it is for all students combined.

Table 2 shows correlation coefficients between training success and the sum of the two MCAT scores as well as this sum plus the OKT score. It can be seen that adding the OKT score to the sum of MCAT scores causes a substantial increase in the validity of the composite score. Table 2 also indicates that both test composites have a very high degree of validity for students in the 150 program, for women and for minority students.



Performance on the Project TALENT Information Test Total (R 190) of Groups of High School Boys (Tested as Seniors in 1960) Who Were Members of Important Educational or Occupational Groups in 1961 and Who Responded to a Follow-Up Questionnaire. Selected groups of the total sample are also shown. (Adapted from Dailey, 1964.)

Figure 1

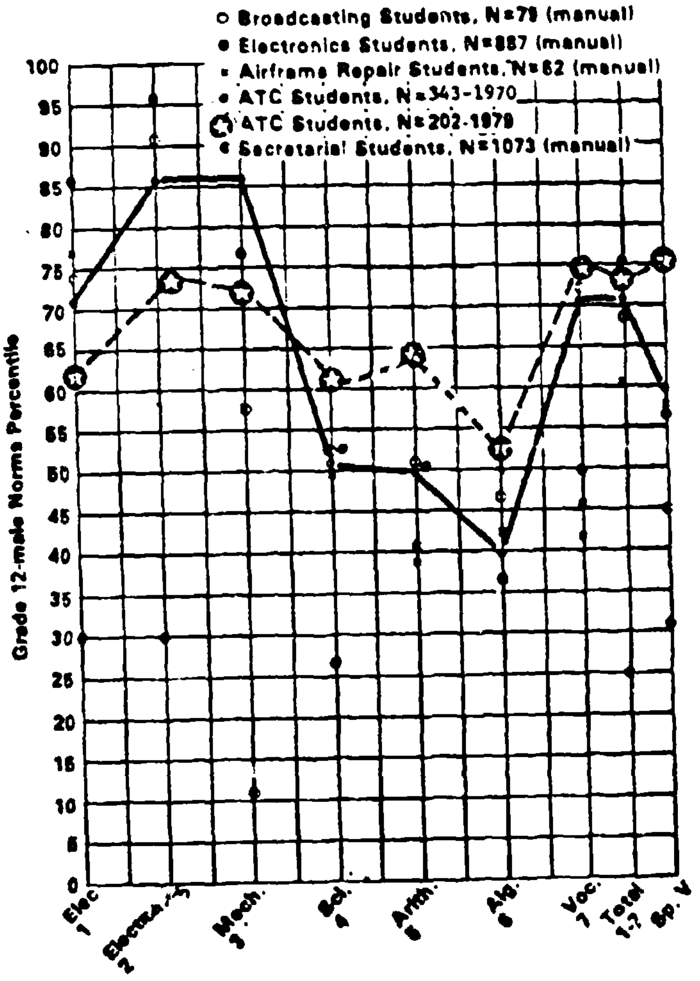

○ Broadcasting Students, N=79 (manual)
● Electronics Students, N=887 (manual)
▪ Airframe Repair Students, N=62 (manual)
• ATC Students, N=343-1970
☆ ATC Students, N=202-1979
◆ Secretarial Students, N=1073 (manual)
Grade 12-male Norms Percentile
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Elec 1
Electr. 2
Mech. 3
Sci. 4
Arith. 5
Alg. 6
Voc. 7
Total 1-7
Sp. V

Figure 2

ATC-ACADEMY EN R+T TESTED

6/20/TO 10/17/78

2/1/79
AAM-500
OAM
JTD

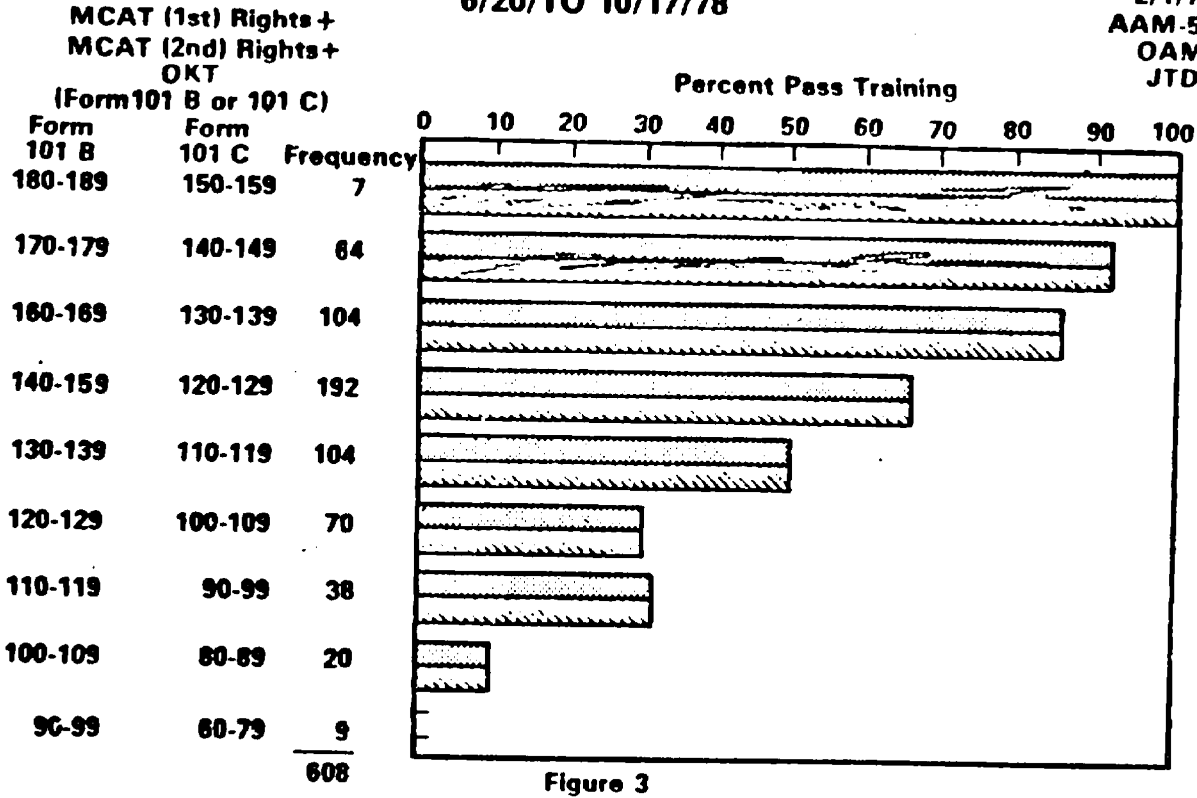


Figure 3
Multiplex Controller Aptitude Test
Occupational Knowledge Test-ATC
Prediction of Training Success-ATC Academy
r_{bis}=.605

ATC-ACADEMY EN R+T TESTED

6/20 TO 10/17/79 – 150 PROGRAM

2/2/79
OAM
AAM-500
JTD

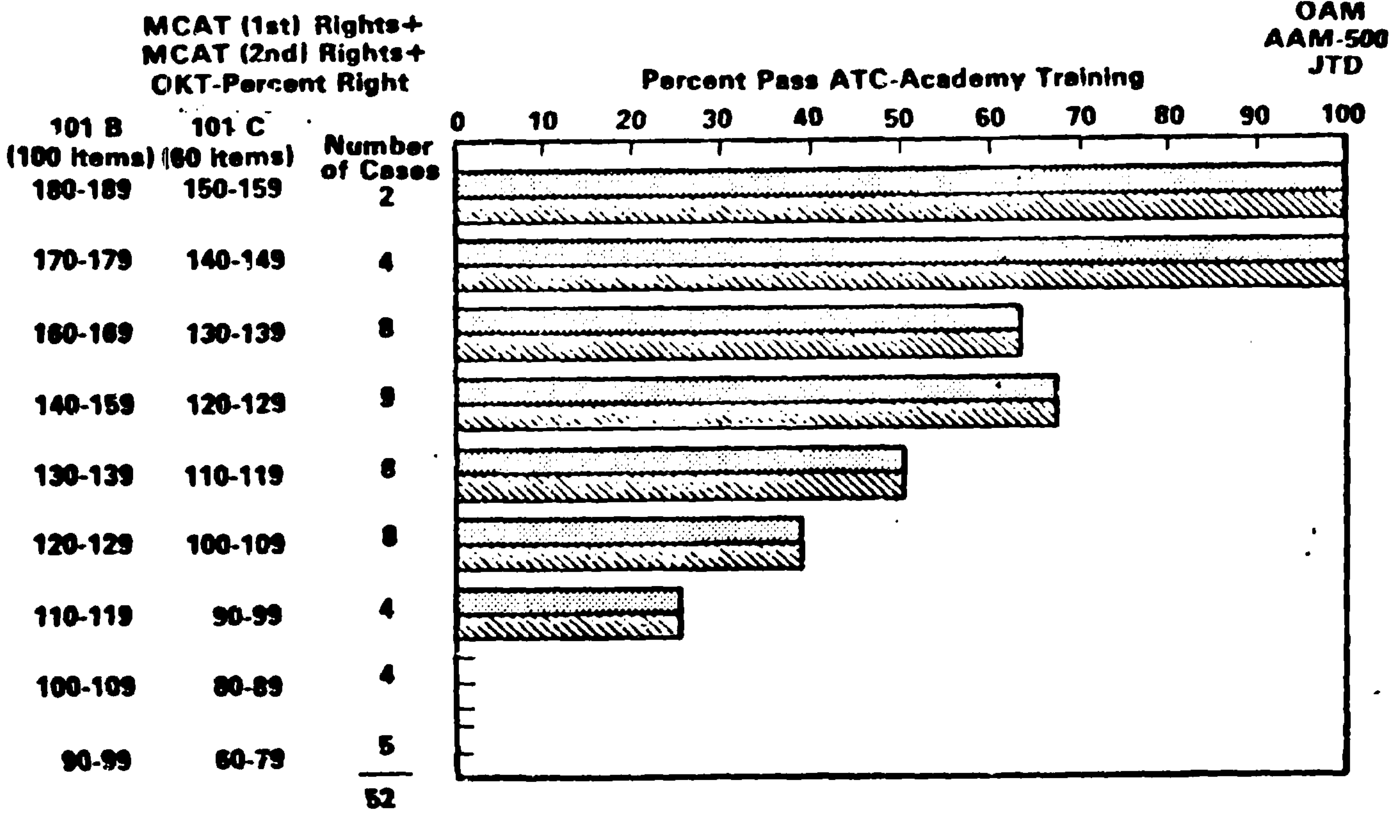


Figure 4

r_{bis}=.715

ATC-ACADEMY EN R+T FEMALE TESTED

6/20/ TO 11/17/78

2/5/79
JDT

MCAT (1st) Rights + MCAT (2nd) Rights + OKT		
101 B (100 items)	101 C (60 items)	Number
180-189	150-159	1
170-179	140-149	3
160-169	130-139	14
140-159	120-129	20
130-139	110-119	23
120-129	100-109	16
110-119	90-99	9
100-109	80-89	7
90-99	60-79	5
		98

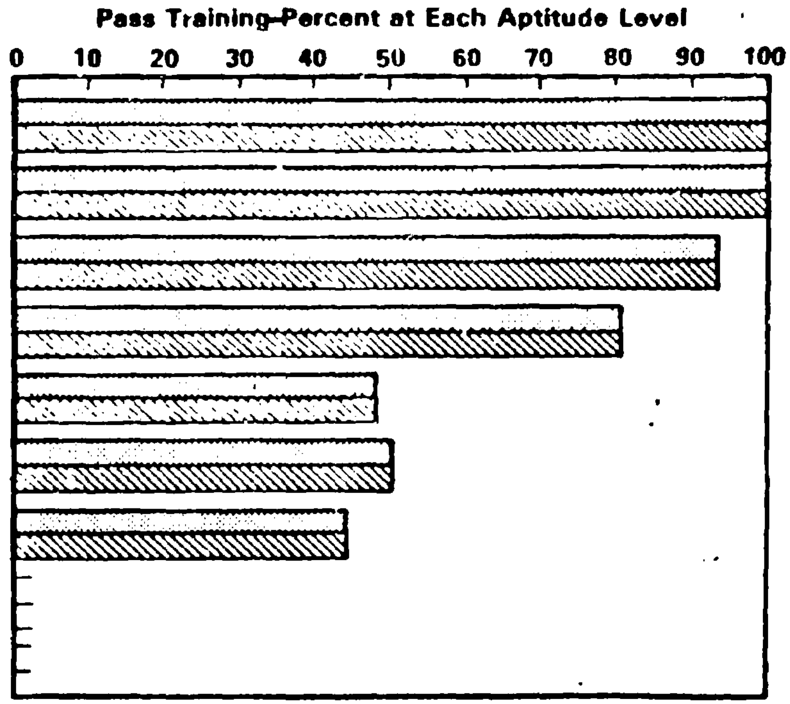


Figure 5
All Women
Aptitude vs. Pass-Fail
r_{bis} = .685

Table 1

AIR TRAFFIC CONTROLLER VS. GENERAL POPULATION

CSC ATC Composite Score	General Population	ATC Trainees 1961-63 (N=893)	ATC Trainees Still on Job-1968 (From '61-'63) (N=501)	ATC Incumbents GS-12 &13 (N=140)
270 +	.04%	1.5	2.6	6.4
250-269	.40%	6.3	9.4	26.4
230-249	2.3%	18.7	24.6	48.6
210-229	9.2%	35.6	44.4	63.6
190-209	24.8%	54.9	65.6	80.0
170-189	49.7%	74.5	84.6	89.3
150-169	74.6%	85.6	93.0	96.5
Below 150	100.0%	100.0	100.0	100.0

Table 2

ATC—Academy En R+T Tested 6/20 to 10/17/78

2/5/79
OAM
AAM–500
JTD

Biserial Correlation with Training Success (Pass vs. Fail or Resign)

Group	Number of Cases	MCAT (1st) Rights + MCAT (2nd) Rights	MCAT (1st) Rights + MCAT (2nd) Rights + OKT (101 B or C or 102 ABCD)	OKT
All Cases 6/20 to 10/17/78	605	.514	.589	
All Men	507	.494	.582	
Men Non–150	481	.477	.607	
Men–150	26	.686	.746	.743
Women–150	30	.876	.783	.215
All 150 6/20 to 10/17/78	55	.785	.784	.451
All W~men	98	.605	.685	
Women Non–150	68	.440	.684	.680
All Minority 6/20 to 10/17	68		.792	
All Cases Tested 6/20 + 8/1	262	.457 (one MCAT r = .381)	.584	.390
All Cases 8/29 to 10/17/78	343	.550	.608	.351
All Cases 6/20 to 11/21/78	763		.5854	
All 150 6/20 to 11/21/78	82		.707	
All Cases				
Tested 11/21/78	90	.468	.510	
Tested 10/24/78	66		.759	
Tested 10/17/78	107	.569	.636	

Prediction of Success for Air Traffic Controllers

by Joseph G. Colmen, Ph.D.

Introduction

In January 1970, a blue-ribbon committee was empaneled by Secretary of Transportation John Volpe to study and report on all facets of employment which affected the career and well-being of the air traffic controller. John J. Corson was appointed chairman of the distinguished group which came to be known as the ATC Career Committee.

The Committee soon observed that "while many categories of employees must possess some of the talents; and while many other jobs impose some of the exacting responsibilities, few combine as many demands upon the individual as does the job of the controller....(In addition)....there is compelling evidence that many controllers work for varying periods of time under great stress. They are confronted with the necessity of making successive life and death decisions within very short time frames--decisions requiring constant standards of perfection." It was little wonder, then, that among the major recommendations made by the Committee, selection research took a prominent position, including reconstructing the written tests, looking at differential placement within the various ATC options (Terminal, Enroute and Flight Service), considering qualifications of education and experience, analyzing the potential discriminatory effects of current selection procedures, and developing more objective, systematic means for evaluating proficiency of controllers on the job.

In that year, FAA contracted with Education and Public Affairs (EPA), for what was to become five years of research into the selection process used with air traffic controllers. While the ATC Career Committee report undoubtedly stimulated the research, a more immediate economic motive existed: an unacceptably high attrition rate of 25% to 40% during the two to five year training period required to reach the full performance or journeyman level. Most of the attrition was attributed to the inability of trainees to acquire and demonstrate the skills and knowledges required to progress satisfactorily through the training. The costs of that attrition were calculated at over four million dollars, to say nothing of the irretrievable loss of "time" occasioned by the need to start a replacement at the beginning of the training cycle. Even earlier identification of those most likely to "wash out" could have enormous economic consequences.

The first of the two major selection research studies undertaken by EPA was initiated in 1970 (Milne and Colmen, 1972); the second in 1975 (Mies, Colmen and Domenech, 1977).

The 1970-71 Study - This research was addressed to three primary questions:

- To what extent it is possible to predict the quality of job performance of a journeyman ATCS from a battery of tests administered at the time of his application.

- To what extent improvement can be achieved by assigning applicants consistently to one of four selected ATC options (FSS, Center, IFR or VFR terminals) and within these, into high or low density (activity) facilities.

- To what extent the measures selected to accomplish the above objectives affect black and white applicants with an equal degree of fairness.

Sample Description

The total sample comprised approximately 800 employees who were either journeymen ATC Specialists or new ATC appointees including, because of the small numbers of those groups, an oversample of black and of women ATC Specialists. Sample selection provided for distribution among the four types of ATC facilities (Centers, IFR and VFR terminals, and Flight Service Stations). Samples for Centers and IFR terminals were further stratified between high and low activity facilities. Racial distribution, including the oversample was 93 percent white and 7 percent black.

Using 15 major cities in the U.S. as hubs, the sample was randomly selected (except for the oversample) from ATC facilities within a 100-mile radius of 15 metropolitan area hub cities. In recognition of earlier CAMI research (Cobb, 1967 and Cobb, Lay, and Bourdet, 1971) showing consistently negative correlations between age of entry into ATC work and later performance as a controller, sampling was controlled to exclude ATC specialists over 36 years of age and to insure that full performance specialists had no less than three nor more than ten years of ATC experience with FAA. Because of the small number of women in the ATC workforce, it was not possible to stratify sample selection on the basis of gender. All subjects, participating voluntarily, were given a battery of paper-pencil tests or forms, in addition to the current CSC test battery. In addition, a subsample of about 260 ATC Specialists who took the paper-pencil battery were also given a series of "psychomotor" tests in a separate testing session at FAA facilities at Oklahoma City. In total, 34 percent of the journeymen took the psychomotor tests together with 32 percent of the new appointees. Selection of this subsample was controlled for region, type of ATC facilities and variance on the confidential supervisory performance evaluation which was used as the criterion measure. Racial distribution of the "psychomotor" subsample was 83 percent white and 17 percent black.

Predictors - To be selected for this study, predictors had to meet four criteria:

- Cover as broadly as possible the range of job/tasks and worker attributes identified with the ATC occupation.
- Not substantially overlap areas already covered by the existing CSC test battery which was also to be administered to the ATC specialists.
- Based on prior research, have potential validity for ATC selection.
- Require no more than eight hours for each part (paper-pencil and psychomotor) to administer.

The following were identified for use as experimental predictors:

- Paper-Pencil Battery
    - Aptitude tests: the current five-part CSC battery (arithmetic reasoning, spatial relations, air traffic problems, following oral directions, abstract reasoning and a composite score); Minimum Coins; and Dial and Table Reading.
    - Knowledge and Interest tests: Dailey Technical and Scholastic Test (electricity, electronics, mechanical information, physical sciences, arithmetic reasoning, elementary algebra and vocabulary; and an ATC General Information Test, comprising items dealing with low-level aeronautical information.
    - "Personality" tests: Ben Graham's Concept-Adjective Test, patterned after the Semantic Differential with three concepts--the job of controller as it is, yourself in the job, and the ideal controller; and the Closure Speed Test which requires the respondent to identify patterns from incomplete outlines.
    - Background Information: A specially constructed biographical inventory.
- Psychomotor Battery
    - Controller Decisions Evaluation (CODE), developed at NAFEC, a filmed version of radar screen air traffic with dynamic or changing patterns, from which the respondent must anticipate aircraft conflictions; Multiple Task Performance Test, developed at CAMI, requiring responses to single and multiple mode stimuli at a large console; Compressed Speech Test, in which the respondent follows instructions given orally (by tape), then responds when the speech is compressed in time; and again when external distractors are introduced concurrently; the Press Test, which provides color coded stimuli with conflicting coding instructions; Hidden Patterns Test, requiring the respondent to find the stimulus pattern embedded in a larger diagram; and Directional Heading, a paper-pencil test of conflicting weathervane directions, distracted by tape driven oral stimuli.

The paper-pencil battery was administered to the total sample group at facilities near the controllers' place of employment. The psychomotor battery was given to a subsample of that group at FAA facilities in Oklahoma City where equipment necessary for administration of the Multiple Task Performance Test was available. All test results and other information on individual ATC specialists was and has been kept confidential by Education and Public Affairs.

Criterion Measure

After examining a number of alternatives, such as System Error Review Board records, peer ratings, and the Dynamic Simulator, a computer driven real-life exercise at NAFEC, a Confidential Supervisory Evaluation form was selected as the criterion against which experimental tests were to be validated. These evaluations were obtained by Education and Public Affairs directly from the ATC specialists' supervisor. They were not reviewed by FAA nor were copies provided to the agency. The evaluation covered the broad areas of performance derived from job analysis, with a number of task behaviors for each area. These included Knowledge, Perception, Comprehension, Memory, Communication, Judgment, Traffic Management Techniques, Performance Work Stress, Interpersonal Skills, and Other Personal Skills. In addition, an "Overall Performance" category consisting of four general task items and a summary evaluation were obtained, the latter being a seven-part rating scale. Two measures were finally derived from the Supervisory Evaluation for use as criteria: the unit weighted sum of 18 selected items drawn from the factor analysis of the instrument, particularly the loadings on the general factor; and item 50, the seven-part overall rating. The r between these was .77. The distribution of the total sample of 597 full performance level specialists on the summary performance evaluation item (50) was: Top 10%, 25.3%; next 15%, 35.6%; next 15% 20.6%; middle 20%, 12.7%; next 15%, 3.7%, next 15%, 1.2%; and bottom 10%, .7%. While somewhat skewed, the distribution was felt to be sufficiently broad to permit reasonable expressions of validity to emerge. It was found that groups which had claimed prior ratable experience were rated higher than those who entered on the basis of test scores alone. While these differences were statistically significant, they were of small practical significance.

New appointees were found to do significantly better on tests than journeymen, possibly because they had come in recently on the basis of tests. Journeymen, on the other hand, did significantly better on performance or psychomotor tests, possibly because they were more job specific.

Descriptive Information

Based on biographic responses from 304 FPL ATC the sample was defined as follows: 98% were men, 2% were women; 96% had prior military edperience, 72% as a controller; 2% claimed experience as a pilot; 48% had taken the CSC test for appointment; 25% attended college, but fewer than one percent had completed it.

Analytical Methodology

Data collected were subjected, as appropriate to the question to be addressed, as follows:

- Multiple regression analysis, to determine how well predictors related to job performance.

- Analysis of variance to test differences between such groups as minority and non-minority, journeymen and new appointees, and persons hired with or without prior aviation-related experience.

- Multiple discriminant function analysis, to evaluate tests in terms of their ability to optimize placement between the various options and activity levels.

Results

In summary, this study concluded that:

1. Capacity of the paper-pencil tests to predict job performance of journeymen ATC specialists produced mixed results. The CSC test was marginal in predicting job performance, as measured by supervisory evaluations, but questions of restriction in range made the results inconclusive. When supplemented by additional experimental paper-pencil tests, gains in predictive capability were small. The psychomotor tests, however, showed consistently significant correlation with supervisory job performance evaluations.

2. By combining paper-pencil and psychomotor tests with most promising validities, it was possible to improve placement success, i.e., assignments to the different ATC options. It was found that accuracy of assignment as between Enroute, IFR, VFR and FSS could be increased from 25% (random probability) to 58%. A paper-pencil battery alone could produce an accuracy level of 40%. With respect to assignment to facility activity levels (restricted to IFR and Enroute only, due to sample sizes), classification accuracy levels as high as 75% to 80% were achievable with the addition of two psychomotor tests, beyond those required for option placement.

3. If relevance to the job is the primary factor in determining test acceptability, the tests proposed met that criterion. They also predicted job performance equally well for black and white members of the sample. Supervisors' job performance ratings were also similarly distributed for both racial groups. However, as a group, blacks consistently scored lower on the tests than whites.

This study provided FAA with valuable information and insights on the problems associated with selection and placement of applicants for ATC work. It further stimulated research within FAA, especially on development of training criterion measures and on development of more job relevant tests, such as MCAT. However, action to change selection and placement procedures was deferred due to a number of considerations, including the complexity of the specialized equipment required for the psychomotor tests, logistical difficulties in test administration and complexity in test scoring and ranking of applicants.

1975-77 Study

Continuing concern with the rate of attrition of ATC trainees led FAA to contract with Education and Public Affairs for a follow-on analysis.

Objectives

The objectives of this research were directed to essentially the same concerns as the previous study (Milne and Colmen, 1972) selection; placement; and, fairness. However, the study design significantly expanded the sample size, structure, and representation; encompassed more criterion measures of ATC job success; and incorporated evaluation of prior aviation-related experience and educational level as predictors of ATC job success.

Longitudinal Analysis of 1971 Experimental Tests

As preliminary to the new study, a longitudinal analysis was conducted to ascertain how well the experimental test results obtained in 1971 related to criterion measures that could be obtained in 1975. In addition, a factor analysis of the original battery of 14 paper-pencil tests was accomplished to permit reduction in battery size on the basis of overlapping of underlying variables. The tests thus selected were subjected to multiple regression analysis, allowing further reduction by identifying the minimum number of tests or test scores which would predict the maximum proportion of variance in the criterion measures.

Of the fourteen predictors administered in 1971, eight were found to warrant further experimentation in the new study. Table 1 shows, by the entry X, those tests which were valid for each of the criterion measures accessible in 1975.

Sample Description - 1975-76 Study

A comprehensive sample design was constructed to define the ATC population to ensure a representative sample for three specific "year of hire" groups. There groups represented three ATC career "stages": (1) New Hires (1976); (2) Developmental ATC Specialists with 2 to 3 years ATC experience with FAA (1973 and 1974); and, (3) ATC Specialists with 2 to 6 years experience at the journeyman (FPL) level (1969 and 1970).

In addition, two other ATC specialist samples were included: an oversample of currently employed women and minority ATC specialists in the same three year-of-hire groups; and a sample of ATC specialists hired during the three time periods sampled, but who had separated from ATC work before reaching journeyman status.

Employees who were in ATC staff or supervisory positions or, except in the FSS option were over 31 years of age at the time they were hired, were excluded, because 31 was then the maximum age for initial employment.

Sample selection for the three primary groups was based on stratified random sampling methods to provide a proportionally representative group of the total constrained ATC universe for each of the four ATC "options" (FSS, VFR, IFR, and ARTCC) with respect to both initial and current option of assignment. Initial, or first option of assignment, and current option of assignment were both important, for two reasons: (1) The progression criterion represented the discrepancy between the two; and (2) certain criterion measures, such as attrition, were considered more highly related to initial assignment, while others, such as supervisory ratings, would be more highly related to current assignment.

Table 2 identifies for the various ATC year group samples, the total number desired and the samples actually obtained. "FPL" refers to full performance level or journeyman; "DEV" refers to developmental or below journeyman level.

Because the number of women and minorities who volunteered was not adequate for analysis by year group or ATC option, the analysis on test fairness was based on combined year and option groups testing 235 women and 321 minorities. The oversample for women and minorities was used only for analysis of the fairness of the ATC "success" predictors.

Predictors

In addition to the predictors derived from the longitudinal analysis of the experimental tests administered in 1971, five other instruments were included:

> CODE (Controller Decision Evaluation) This test consisted of three film versions of a computer simulation of moving air traffic patterns appearing on a radar scope. Initially these were converted to slide projector presentation to eliminate the need for movie projector equipment and to simplify both the response recording and scoring, but group administration clearly pointed up the practical problems of using film or slide projection equipment in test administration. Consequently, a paper-pencil version was developed by Dailey and Pickrel which incorporated structured measures of abilities to identify potential conflicts of aircraft as well as the traditional kinds of aptitudes within an air traffic control context. The resulting tests were entitled the Multiplex Controller Aptitude Test (MCAT).

Arithmetic Reasoning

In the prior research, the arithmetic reasoning test was part of the Dailey Technical and Scholastic Test. Since it was commercially published, it was not acceptable to the CSC for competitive examination. Instead, a similar test developed by the Army Air Force was substituted.

Pre-Employment Experience Questionnaire (PEQ)

To obtain specific data from participating ATC specialists on various kinds of pre-FAA experience and education, a questionnaire was developed from the CSC Rating Guide elements used as a basis for granting additional credit in the ATC employment selection process.

ATC Occupational Knowledge Test (OKT)

Administered only to trainees in the sample. Because this test was designed to be "job-knowledge specific" it was not included to evaluate its use in screening ATC job applicants for employment eligibility, but rather to measure the "quality" of prior experience as a potential improved basis for granting additional credit for experience in the selection process in place of the existing CSC Rating Guide.

Sixteen Personality Factor Questionnaire (16 PF)

The 16 PF Questionnaire had been administered as part of the medical qualification process to all entering ATC Specialists. Designed to measure important personality characteristics not otherwise measured, it was included here to determine its possible utility for selection or placement purposes.

While it was the intent to obtain CSC test scores for participating ATC specialists from existing records, this did not prove to be feasible for a large number of employees in the sample. Consequently, analysis of the experimental tests in relation to the existing CSC test battery as predictors of ATC success was not possible.

Pre-employment experience and education information was obtained on specially designed forms from ATC Specialists at the time they agree to participate in the study. Test data were administered to the new hires at the FAA Academy on their first day of attendance. For those ATC Specialists assigned to facilities, the tests were given by FAA test administrators trained by Education and Public Affairs.

Criteria of ATC Success

To determine how valid experimental tests, prior experience and education were as predictors of ATC success, four criterion measures were used in the study; these were then combined into a single "aggregate" measure of ATC success.

Training Performance-- measured by scores received on the ATC Laboratory Problems and the Controller Skills Test during initial ATC training at the FAA Academy. These scores were selected for their high level of job relevance, requiring students to demonstrate operational application of academic knowledge.

On-the-job Performance-- measured by confidential job/task assessments prepared by the employee's supervisor. It included 54 questions on ATC job tasks and four general questions on quality of job performance. Responses to a seven-point "overall" ranking scale were selected as the measure for on-the-job performance.

Progression-- measured by the ATC "option" to which the ATCS was assigned to on January 1, 1976. The assumption behind the progression criterion was that employees who advance to more complex options within the ATC profession were more desirable, having maximized their training and operated at a higher level of ability. Those who were at the same or lower levels of complexity were deemed less successful.

Attrition-- measured by whether or not ATC Specialists hired during the year groups sampled for this study were still employed in ATC work. Those still employed as ATC specialists were assigned a "high" score; those separated were assigned a "low" score.

Aggregate Criterion of ATC Success-- constructed from the four individual criterion measures, (training, on-the-job performance, progression and attrition) yielding a five point scale value for ATC "success."

Descriptive Information

The most marked changes between the three "year-of-hire" groups (and the 1972 sample) were in education level, military service and pilot experience. Table 3 compares the various sample groups on selected variables. It is especially interesting to note that the proportion of ATC specialists with a college degree had risen over the years: 13% with a degree were in the 1969-70 and 23% in 1973-74: 28% of the trainees in the 1976 sample had college degrees.

Scheme for Analysis of Data

In brief, the validation process followed five steps:

1. Analyses of each of the smallest homogeneous samples (i.e. each ATC option and all ATC options combined within each year group against each of the four individual ATC success criterion separately). This resulted in 118 separate analyses.

2. Predictors selected from Step 1 (based on validity coefficients and significance levels) for each ATC option and all ATC options combined across all year groups were then analysed to identify the best overall set of predictors for each of the four individual criterion measures. This resulted in 17 separate analyses.

3. Results of each of the four individual criterion measures were then examined for each ATC option and all options combined to determine the best set of predictors across criterion measures. This resulted in 17 separate analyses.

4. The final set of predictors from step 3 was then validated against the aggregate ATC success criterion leading to derivation of weighted test and experience scores. This involved five separate analyses.

5. The weighted test battery and experience scores based on step 4 were then validated against the four single criterion measures. This involved four separate analyses.

In the case of the pre-employment questionnaire, the test validation methodology was suplemented to establish an empirical scoring key for items on the basis of their relationship to the various criterion measures, although two additional scales were formulated: one related to the CSC rating scale for experience and education; and another based on an *a priori* conceptualization.

The methodology to evaluate the fairness of the experimental test battery, prior experience and the occupational knowledge test was directed to determining differential validity in accordance with the Uniform Guidelines on Employee Selection Procedures (15) which state:

> "When members of one racial, ethnic or sex group characteristically obtain lower scores on a selection procedure than members of another group, and the differences are not reflected in differences in measures of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores."

Results- Selection

Two of the experimental tests-- Multiplex Controller Aptitude (MCAT), and Directional Headings-- predicted the ATC success criteria established for the study at statistically significant levels of confidence for all ATC options combined and for the three primary year groups sampled (1969-70; 1973-74; and 1976). Two other experimental tests-- Arithmetic Reasoning and Dial Reading-- either did not predict the ATC success criteria or did not add appreciably to the prediction values obtained from MCAT and Directional Headings.

The weighted battery predicted the aggregate criterion for all options and all years combined at R = .265, $p \leq .01$. Pre-employment aviation related experience and the ATC Occupational Knowledge Test, while not intended for use in determining initial appointment eligibility, predicted ATC success at statistically significant levels of confidence increased the validity coefficients obtained with the experimental tests. The empirically keyed scale predicted the aggregate criterion for all options and all years combined at R = .235, $p \leq .01$. Addition of the scale to the weighted battery increased validity to R = .341, $p \leq .01$. For IFR and ARTCC alone, composite validities jumped to .426 and .372 respectively. The validities derived from the analysis of combined predictors against the Aggregate ATC Success Criterion by ATC option for all year groups combined are provided in Table 4. Note: The weighted test battery, PEQ and OKT combined predicted the training performance criterion at R = .561 for Terminal trainees and .516 for Enroute trainees, both significant at $p \leq .01$. Level of education prior to FAA employment did not predict ATC success. Other experimental instruments-- 16 Personality Factor Questionnaire, Concept Adjective and Biographical Information Questionnaire-- did not add appreciably to the predictive capability of the experimental test battery, preemployment experience and the ATC Occupational Knowledge Test. The weighted experimental test battery did differentiate between the Terminal (IFR, VFR) and ARTCC options. Average scores for ATC specialists in FSS, Terminals and ARTCCs were different at statistically significant levels of confidence. The average FSS score was lowest and ARTCC highest. Table 5 provides the comparative mean scores for the various ATC option and year groups. Analysis of variance shows that the mean differences between FSS, Terminal (VFR, IFR) and ARTCC are significant at the 1% level for the 1969-70 and 1973-74 groups. The 1976 ATC group difference between Terminal and ARTCC is significant at the 5% level.

Test Fairness

Women as a group scored lower on each of the predictors and on the aggregate criterion of ATC success. In each case, these differences were significant at the 1% level of confidence (Table 6). Validities by sex of the predictors against the aggregate ATC success criterion are provided in Table 7. The fact that few of the 229 women in the prior experience sample had aviation-related experience. May account for the absence of validity of the FEQ for women. Comparable analyses were made between minority and non-minority groups and between non-minority and blacks. The results with respect to the test battery, prior experience and the ATC occupational knowledge test are provided in Tables 8, 9 and 10 respectively Minorities as a group, and blacks as a subgroup of minorities, scored lower on each of the predictors and on the aggregate ATC success criterion. In each case these differences were significant at the 1% level of confidence.

Validities of each of the predictors with the aggregate success criterion for minorities and non-minorities are provided in Table 11. When blacks were analysed separately from all other minorities, the validity of prior experience and the occupational knowledge test was sustained (r =.259 and .256, both significant at the 1% level of confidence). Validity of the weighted test battery was .120 which did not reach the 5% level of confidence (p =.075). However, the difference in the validity for non-minorities and blacks was not statistically significant.

Conclusions

Taylor-Russell tables were used to estimate the proportion of satisfactory employees that could be expected from introduction of the new predictors over present methods. Based on the weighted test battery validity of .26; a selection ratio of .30; and estimates that 25% of those who leave by attrition plus 10% who received supervisory evaluations in the lowest performance categories (total of 35%) represent less than desired levels of satisfactoriness, use of the new battery should reduce that total to 25% with 10% more satisfactory employees, an improvement of 15% (10/65) in the proportion of satisfactory employees could be expected.

The results of the study supported the conclusion that several predictors were of sufficient promise that they should be further developed for operational use in selection of applicants for ATC positions. This work has been ongoing. An important aspect of that work has been investigating the effect an restriction of range, by comparing scores on the experimental battery for new applicants applying through CSC channels with those of trainees and later, controllers on the job.

References

Air Traffic Controller Committee, "The Career of the Air Traffic Controller; A Course of Action, DOT, January 1970

Cobb, B. B., The relationships between chronological age, length of experience and job performance ratings of air route traffic control specialists, Federal Aviation Administration, Civil Aeromedical Institute, Oklahoma City, Oklahoma. (FAA-AM-67-1, June 1967)

Cobb, B. B., Lay, C. D., Bourdet, N. M., The relationship between chronological age and aptitude test measures of advanced level air traffic control trainees, Federal Aviation Administration, Civil Aeromedical Institute, Oklahoma City, Oklahoma. (FAA-AM-71-36, July 1971)

Colmen, J. Validity of the Cattell 16 Personality Factor Questionnaire and Other "Non-Cognitive" Tests for Selection and Placement of Air Traffic Control Specialists. Education and Public Affairs Inc., Federal Aviation Administration Contract DOT FA-75WA-3646, May 8, 1977.

Mies, J., J. G. Colmen and O. Domenech. Predicting Success of Applicants for Positions as Air Traffic Control Specialists in the Air Traffic Service. Education and Public Affairs Inc., Contract DOT FA-75WA-3646, May 8, 1977.

Milne, A. M. and J. C. Colmen. Selection of Air Traffic Controllers for Federal Aviation Administration. Education and Public Affairs Inc., January 1972.

TABLE 1. Predictor and Criterion Measures of ATC "Success" (1975) for Controllers Participating in 1971 ATC Research

	Separation	Progression + Attrition		Present Option		1971 Supv. Assess.	Sup/Staff Position
	DEV	DEV	FPL	DEV	FPL	FPL***	FPL***
CODE	X	X	X	X	X	X	-
Dial Reading	X	X	-	-	-	-	-
Dir. Heading	X	-	-	-	-	X	-
Air Traffic Prob.*	-	-	-	X	X	X	-
Arith. Reasoning**	-	-	-	-	X	X	-
ATC General Info.	X	-	-	-	-	-	-
Concept Adjective	X	X	-	X	-	X	-
Biographical Info.	X	X	-	X	X	X	X

*Air Traffic Problems CSC Test No. 540

**Arithmetic Reasoning was Part 5 of the Dailey Technical and Scholastic Test (TST)

***Supervisory assessment criterion data available for 1971 FPL ATCS only. No new 1971 appointees progressed to ATC supervisory or staff positions by 1975.

TABLE 2. ATC SPECIALIST BY YEAR HIRED, SAMPLE GROUPS AND ATC CAREER STATUS

Year Hired	ATC Sample Year Groups	ATC Career Status	Number Invited	Desired Sample	Obtained Sample	Percent of Desired Sample
1969-70	(1) Employed ATCS	FPL	1344	800	754	(94%)
	(2) ATCS Oversample	FPL	151	200	31	(6%)
	(3) Separated ATCS	DEV	---	---	362	--
1973-74	(1) Employed ATCS	DEV	1127	800	740	(93%)
	(2) ATCS Oversample	DEV	258	200	72	(36%)
	(3) Separated ATCS	DEV	---	---	166	---
1976	(1) Employed ATCS	New Hires (DEV)	610	610	590	(97%)
1971	(1) Employed ATCS	FPL	480	480	270	(56%)
	(2) Separated ATCS	FPL/DEV	---	---	74	---

TABLE 3. EDUCATION AND EXPERIENCE LEVELS

	Year Hired as ATCS			
	1969 1970 (N=659)	1973 1974 (N=661)	1969-70 1973-74 Oversample (N=103)	1976 (N=592)
High School	34%	24%	25%	16%
Some Senior College	53%	53%	59%	56%
College Degree(s)	13%	23%	16%	28%
Military Service	75%	74%	56%	71%
ATC Experience	—	—	—	—
IFR	35%	39%	27%	32%
VFR	39%	38%	26%	37%
Pilot	30%	34%	11%	33%

TABLE 4. VALIDITIES OF COMBINED PREDICTORS AGAINST THE AGGREGATE ATC "SUCCESS" CRITERION (BY ATC OPTION -- ALL YEARS COMBINED)

ATC Option	Weighted Test Battery Scores		Weighted Test Battery Plus PEQ		Weighted Test Battery and PEQ Plus OKT	
	N	r	df	R	df	R
FSS	196	.23**	193	.26**	192	.26**
VFR	479	.26**	474	.44**	423	.45**
IFR	499	.26**	494	.39**	443	.43**
ARTCC	445	.30**	425	.32**	388	.37**
All Options	1309	.26**	1287	.32**	1205	.34**

** = p≤.01

TABLE 5. MEAN WEIGHTED TEST BATTERY SCORES BY ATC OPTION

ATC Option	1969-70 & 1973-74 ATC HIRES				1976 NEW ATC HIRES (ACADEMY TRAINEES)		
	N	Mean	SD*		N	Mean	SD
FSS	196	2274	402		---	---	---
VFR	170	2446	358	(TERM) ----	310	2420	350
IFR	189	2500	373				
ARTCC	182	2643	366		263	2479	384

TABLE 6. - MEANS, STANDARD DEVIATIONS AND t-TEST RESULTS FOR PREDICTORS AND THE AGGREGATE ATC "SUCCESS" CRITERION

	Predictors									Criterion		
	Weighted Test Battery			PEQ			OKT			Aggregate ATC "Success"		
Sex	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Men	1397	3445	380	2736	3.7	2.9	1318	76.6	12.3	1254	3.1	1.4
Women	171	2345	480	235	.8	1.6	158	64.6	16.3	158	2.6	1.4
	(t=3.14;p≤.01)			(t=15.02;p≤.01)			(t=11.21;p≤.01)			(t=4.56; p≤.01)		

TABLE 7. - VALIDITIES OF THE WEIGHTED TEST BATTERY, PEQ AND OKT AGAINST THE AGGREGATE ATC "SUCCESS" CRITERION BY SEX

(*=$p \leq .05$; **=$p \leq .01$)

	Weighted Test Battery		Prior Experience (PEQ)		ATC Occupational Knowledge (OKT)	
Sex	N	r	N	r	N	r
Men	1386	.23**	2721	.21**	1308	.22**
Women	265	.19**	229	-.04	154	.14*

TABLE 8. - TEST SAMPLE

	PREDICTOR Weighted Test Battery			CRITERION Aggregate ATC "Success"		
	N	Mean	SD	N	Mean	SD
Non-Minorities	1323	2477	371	1308	3.2	1.4
All Minorities	245	2204	429	243	2.7	1.4
	(t=10.30;$p \leq .01$)			(t=5.09; $p \leq .01$)		
Non-Minorities	1323	2477	371	1308	3.2	1.4
Blacks	145	2074	407	144	2.4	1.3
	(t=12.30; $p \leq .01$)			(t=6.31; $p \leq .01$)		

TABLE 9. - PRIOR EXPERIENCE SAMPLE

	PREDICTOR Prior Aviation Experience (PEQ)			CRITERION Aggregate ATC "Success"		
	N	Mean	SD	N	Mean	SD
Non-Minorities	2115	3.9	2.9	2097	3.2	1.4
All Minorities	321	2.4	3.0	318	2.6	1.4
	(t=8.48; $p \leq .01$)			(t=7.15;$p \leq .01$)		
Non-Minorities	2115	3.9	2.9	2097	3.2	1.4
Blacks	194	1.7	2.8	193	2.4	1.3
	(t=10.27; $p \leq .01$)			(t=8.39; $p \leq .01$)		

TABLE 10. - ATC OCCUPATIONAL KNOWLEDGE SAMPLE

	PREDICTOR			CRITERION		
	ATC Occupational Knowledge Test			Aggregate ATC "Success"		
	N	Mean	SD	N	Mean	SD
Non-Minorities	1247	76.5	12.8	1235	3.2	1.4
All Minorities	229	69.1	14.7	227	2.7	1.4
	(t=7.85; p≤.01)			(t=4.86; p≤.01)		
Non-Minorities	1247	76.5	12.8	1235	3.2	1.4
Blacks	134	66.9	15.2	133	2.8	1.3
	(t=8.12; p≤.01)			(t=3.72; p≤.01)		

TABLE 11. VALIDITIES OF THE WEIGHTED TEST BATTERY, PRIOR EXPERIENCE AND ATC OCCUPATIONAL KNOWLEDGE AGAINST THE AGGREGATE ATC "SUCCESS" CRITERION

(**p≤.01)

	Weighte Test Battery		Prior Experience		ATC Occupational Knowledge	
	N	r	N	r	N	r
Non-Minority	1308	.20**	2097	.15**	1235	.20**
All Minorities	243	.22**	318	.25**	227	.27**

Criterion Development for Air Traffic Controller Training

by John T. Dailey, Ph.D. and James I. Moore

Introduction

Attrition and screening for selection of Air Traffic Controllers in the Federal Aviation Administration has long been the subject of much discussion, as well as considerable research and analysis. Based on the analysis a major change in selecting candidates for controller duties is being implemented.

In June 1974 a short term committee was established by the FAA to review selection and retention of controllers. Unable to provide conclusive recommendations, the committee recommended that a task force be established to further study the problem. The newly formed task force recommended that the basic impetus for review should be (1) the attrition rate for trainee controllers; (2) the disservice to individuals who do not qualify as journeymen; and (3) the cost avoidance which would accrue through a more selective appointing process.

In January 1976 the FAA implemented a new training program in which one of the prime elements is mandatory participation by all controller trainees. Everyone hired as a trainee controller must satisfactorily progress through the entire training program to be certified as a full performance level controller. Trainees who fail any phase of training (there are six in the Terminal option and thirteen in Enroute) are immediately removed from the air traffic control career.

The training, which is conducted in classrooms and laboratories, is constantly assessed. The assessment instruments are validated and are reliable indicators of the trainee's ability to perform at the full performance level.

A Congressional Committee, The House Government Affairs Committee released a report recommending that the FAA improve its selection and training procedures for air traffic control specialists because of the high costs and unacceptably high number of controllers attriting from the system. They suggested that the FAA review the criterion and selection devices used by the Civil Service Commission and develop a test battery that would more accurately reveal whether a candidate for controller will succeed. The Committee further recommended that criteria for screening and eliminating unsuccessful students be established and used at the FAA Academy, as well as later in the training program, to ensure that potentially unsuccessful controllers are eliminated early in the training process.

The FAA is responding to these recommendations. The Office of Aviation Medicine has developed new selection tests for use by the Office of Personnel Management. These tests are to replace a part of the out-dated ATCS battery. All new tests developed by the Office of Aviation Medicine, in the FAA, adhere to the American Psychological Association's Standards for Educational and Psychological Tests. Currently two new tests have been developed and adopted for controller candidates. These are the ATC Occupational Knowledge Test and the Multiplex Controller Aptitude Test.

The Occupational Knowledge Test is a subject element, paper and pencil, machine scorable type test having 100 multiple choice items. The test is designed to measure the subject knowledge determined to be important to performance of the controller's job.

The Multiplex Controller Aptitude Test was designed as a highly job-related test of the skills used by the air traffic controller. Over eight types of items have been designed into the test to measure different aptitudes. In effect, the test teaches the subject simplified rules for a very complex "game", then evaluates how well the candidate can play the "game".

As a necessary part of the newly implemented controller training program, a new test was developed, the Controller Skills Test (CST). The test was designed to measure the application of knowledge and skills during the first months of Air Traffic Control Specialist training. Three basic elements for evaluation are distributed within the test: (1) Application of separation standards - students must respond to situations expressed by flight strips and or controller charts, (2) Responding to or forwarding information received which pertains to coordination, with other controllers, and (3) Other items such as Board Management, Timeliness of Actions and Phraseology in developing . the criteria for the CST, field data was gathered. Subjects from six Terminals and four Enroute Centers, a total number of 226, were administered the prototype CST along with a battery of tests from the early FAA Academy Training Program. The subjects taking the test battery at the various FAA facilities ranged from 0 to 200 plus months of air traffic experience.

On the non-radar subject matter of the prototype CST, it took a considerable time for ATC's to achieve mastery. Few, even with extensive ATC experience, made much of a showing on either test until their second or third year on the job. By then, however, the facility controllers did relatively well on the test. This indicates that the CST is relatively independent of prior experience and measured skills learned on the job. On the written test taken from the Academy battery, the peak of performance was reached after about 12 to 21 months, and the rate of mastery seems to begin to fall off after 22 to 39 months' experience on the job. On the CST, the peak of plateau was reached by those with 22 to 39 months experience on the job, and the fall off did not begin until 40 to 69 months' experience.

This would indicate that, to some significant extent, the CST does measure the application of the subject matter, rather than mere knowledge of it. The test measures the ability to handle considerable chunks of information and to make ATC-type judgments regarding it. Thus the CST would represent an important addition to the written test and criterion problem evaluations.

In the early phases of training, (non-radar control phase) while at the FAA Academy, each student takes six laboratory criterion problems which are administered and evaluated by six separate instructors. Each problem takes about one hour and involves a simulated non-radar control task to be performed under very realistic conditions. Each problem is scored for points by the instructor, according to various types of errors and deficiencies made by the student. Additionally, each student's performance is evaluated by the instructor on a rating scale from 40 to 100. After the student successfully completes the six laboratory problems and the instructor's ratings, he is administered the CST. The CST is a 100 item test having a one hour completion requirement. Thus the student is required to make ATC judgments and answer multiple-choice items under time pressure as will be required on the job.

The distributions of scores on several of the assessment measures in the Non-Radar Phase are shown in Table 1. It can be seen from Table 1 that the point scores on the Criterion Problems tend to be high, with the most frequency score group being from 95 to 100 percent. The scores run somewhat lower on the instructor ratings and on the Controller Skills Test. It can be seen from these data that the reason for the absence of attrition in the first class was primarily the student's extremely good performance on the practical lab work as measured by the point scores. While an occasional student might fail one problem, almost no one consistently failed enough of them to have a failing average on the sum of the four problems.

In evaluating the student assessment measure, it is important to obtain an estimate of the consistency or reliability of each of the assessment measures. For this purpose, an analysis was made of the interrelationships of each measure on each of the four Criterion Problems and also the Controller Skills Tests. There were three scores on each problem: point score, instructor assessment, and the average of the two. There were 13 measures in all, and the intercorrelations among these are shown in Tables 2 and 3 for Enroute and Terminal Training.

It is possible to obtain an estimate of the reliability of a single problem (the estimated extent to which it correlates with itself) by computing its average correlation with the other three problems. It can be seen that the reliability of a single Criterion Problem is too low to warrant use by itself for attrition. From the Spearman-Brown prophesy formula, *the reliability for sums of two or more Criterion Problem scores can be computed. (This estimate of reliability is the estimated correlation between the sum of two problems and the sum of two other similar problems). The data in Table 2 indicates that, for the Enroute Course, the instructor's evaluations and average scores are more reliable than the point scores. It is estimated that the reliability of the sum of average scores for all four problems is .55. This is not reliable enough to use separately. If there were two more problems, however, the reliability for all six problems would be .64 and could be used separately if desired. The instructor's evaluation makes a valuable contribution to supplementing and strengthening the point score. Tables 4 to 9 show how performance on one problem compares with that on another. It can be seen that the instructor ratings are more consistent than are the point scores. The Controller Skills Test correlates with the problem averages almost as well as they correlate with themselves. This indicates that CST is measuring substantially the same things as the Criterion Problems. When combined with the four problems, a total is produced which has adequate reliability for separate use in attrition. The correlation between Problems 10 and 10A is only .18, but is .33 between Problems 15 and 15A. It is believed that 10 and 10A are given before learning has progressed to the point where stable performance can be exhibited. Another week of training and two more problems could yield an average score on the last four problems with an estimate reliability of .66. The number of problems was increased to six.

The data in Table 3 for the Terminal Class are somewhat different from those for the Enroute group. Problems 1 and 2 correlate .34 and Problems 3 and 4 correlate .25. Here the first problems are fully as reliable as the later problems and the estimated reliability for the average of four problems is .61 and would be .70 for six problems. This indicates that the average of the present four problems could be used separately for attrition. For the Terminal group, the Controller Skills Test correlates substantially lower with the Criterion Problems than they correlate with each other. This indicates that the CST is measuring something different from the problems or the academic tests. It does measure aspects of the Lab Phase Training as well as similar training in the field. It seems to measure those parts of the Non-Radar Lab Phase that are least likely to be known or easily learned on the basis of prior experience and that have to be learned after entry to FAA.

Pass-fail in the ATC training course is only on the basis of a weighted composite of all performance measures and tests for the entire course. The major weight is given to measures on how well the student performs the lab problems. These six problems are given and scored independently by six different instructors, thus minimizing the subjectivity factor. The lab performance is assessed in three different ways (objective point score, instructor evaluation, and objective Controller Skills Test). Accordingly the pass-fail criterion has a high degree of reliability and validity, thus making possible the extremely high validity levels found for new, job-related selection tests.

*DuBois, Phillip H. "An Introduction to Psychological Statistics," Harper & Row, New York, 1965, p. 392

TABLE 1

TEST RESULTS IN NON-RADAR CONTROL PHASE IN ATC ACADEMY NEW AT PROGRAM 4/16/76

ENROUTE, CLASS 1

Composite Scores Phase III

95 - 100	2
94 - 90	32
85 - 89	43
80 - 84	15
75 - 79	2
70 - 74	1
Below 70	0

ENROUTE, CLASS 1, Controllers Skills Test - Low 66%
High 98%
Mean 85%

95 - 99	7
90 - 94	14
85 - 89	27
80 - 84	25
75 - 79	18
70 - 74	1
65 - 69	3

Nothing below these ranges listed above.

TERMINAL, CLASS 1

Controllers Skills Test - Low 52%
High 92%

90 - 94	2
85 - 89	15
80 - 84	20
75 - 79	22
70 - 74	31
65 - 69	8
60 - 64	4
55 - 59	4
50 - 54	2

TABLE 1 - Continued

TERMINAL, CLASS 1, PHASE IV

Composite Scores

95 - +	1
90 - 94	6
85 - 89	33
80 - 84	46
75 - 79	19
70 - 74	3
below 70	0

ENROUTE, CLASS 1

Problem 15

	Points	Rating Scale
95 - +	51	13
90 - 94	12	21
85 - 89	3	19
80 - 84	11	27
75 - 79	11	7
70 - 74	2	4
60 - 69	3	3
59 & below	2	1

Problem 15A

	Points	Rating Scale
95 - +	34	11
90 - 94	19	19
85 - 89	2	23
80 - 84	15	21
75 - 79	13	11
70 - 74	4	5
60 - 69	4	4
59 & Below	4	1

TERMINAL, CLASS 1

Problem 1

	Points	Rating Scale
95 - +	30	8
90 - 94	22	14
85 - 89	18	22
80 - 84	14	26
75 - 79	15	20
70 - 74	6	14
65 - 69	2	3
60 - 64		
55 - 59	1	1
50 - 54		
45 - 49		
40 - 44		

Problem 2

	Points	Rating Scale
95 - +	32	13
90 - 94	20	12
85 - 89	25	19
80 - 84	14	38
75 - 79	7	15
70 - 74	4	5
65 - 69	1	6
60 - 64	3	2
55 - 59	1	
50 - 54		
45 - 49	1	
40 - 44		

TABLE 2

N = 94

ENROUTE ATC ACADEMY

4/16/76

	Problem 10			Problem 10A			Problem 15			Problem 15A				
1	2	3	4	5	6	7	8	9	10	11	12	13	M	S.D.
1. Point Score	.647	.930	.106	.218	.186	.023	.150	.088	.225	.369	.309	.236	82.2	14.7
2. Rating		.870	.053	.148	.120	.208	.358	.312	.201	.421	.318	.346	82.7	10.5
3. Total Average			.102	.202	.179	.113	.246	.193	.245	.437	.353	.299	82.6	11.5
4. Point Score				.628	.933	.097	.167	.139	.170	.215	.212	.070	87.6	12.8
5. Rating					.849	.033	.161	.088	.030	.128	.082	.180	83.4	8.2
6. Average						.089	.169	.131	.143	.215	.195	.121	85.4	9.5
7. Point Score							.561	.913	.207	.324	.281	.113	90.0	10.8
8. Rating								.847	.181	.456	.324	.187	85.5	8.2
9. Average									.212	.427	.331	.163	87.8	8.5
10. Point Score										.701	.945	.089	86.6	12.0
11. Rating											.894	.179	84.8	8.5
12. Average												.130	87.7	9.5
13. CST													85.2	6.7

TABLE 3

N - 108

TERMINAL ATC ACADEMY

	Problem 1			Problem 2			Problem 3			Problem 4			
1	2	3	4	5	6	7	8	9	10	11	12	13	M	S.D.
1. Point Score	.602	.905	.278	.252	.302	.254	.274	.284	.232	.121	.201	.075	87.3	9.3
2. Rating		.877	.183	.343	.283	.192	.306	.264	.174	.130	.165	.003	82.2	8.1
3. Total Average			.262	.326	.327	.244	.322	.301	.218	.143	.200	.050	84.7	7.7
4. Point Score				.531	.901	.310	.360	.363	.148	.169	.179	.059	87.5	10.1
5. Rating					.842	.279	.414	.375	.121	.213	.183	.008	83.3	8.1
6. Average						.334	.441	.420	.152	.208	.201	.039	85.4	8.0
7. Point Score							.677	.937	.120	.250	.207	.159	84.6	13.4
8. Rating								.891	.214	.251	.262	.070	83.6	10.2
9. Point Score									.176	.277	.254	.129	84.1	10.9
10. Point Score										.532	.897	.109	86.6	11.7
11. Rating											.843	.066	82.9	9.5
12. Average												.099	84.8	9.4
13. CST													75.1	9.2

TABLE 4

ATC ACADEMY ENROUTE

PROBLEM POINT SCORE

				PROBLEM 10A		
Problem 10	40-49	50-59	60-69	70-79	80-89	90-100
90-100	2		1	7	3	29
80-89			3	1	2	7
70-79				4	4	12
60-69		1		2	1	6
50-59				2	1	3
40-49				1		

18 failed Problem 10 (16 of these passed 10A)
9 failed Problem 10A (7 of these passed 10)
2 failed both problems
8 failed an average of two problems

TABLE 5

ATC ACADEMY - ENROUTE

INSTRUCTOR RATING

Problem 10						Problem 10A
	40-49	50-59	60-69	70-79	80-89	90-100
90-100			1	1	14	12
80-89		1	2	7	24	9
70-79				2	7	2
60-69			1		5	
50-59		-		1	1	
40-49					1	

9 failed Problem 10 (8 of these passed 10A)

7 failed Problem 10A (6 passed 10)

1 failed both problems

4 failed an average of 10 and 10A

TABLE 6

ATC ACADEMY - ENROUTE

PROBLEM AVERAGE (POINTS PLUS RATING)

Problem 10	Problem 10A					
	40-49	50-59	60-69	70-79	80-89	90-100
90-100			2	5	7	20
80-89			2	3	6	15
70-79			1	6	6	8
60-69				2	5	1
50-59					2	
40-49					1	

11 failed Problem 10A (all of these passed 10A)
7 failed Problem 10A (all of these passed 10)
0 failed both problems
5 failed an average of 10 and 10A

TABLE 7

ATC ACADEMY - TERMINAL

PROBLEM POINT SCORE

Problem 1	Problem 2					
	40-49	50-59	60-69	70-79	80-89	90-100
90-100				7	13	32
80-89	1		2		18	11
70-79		1	1	3	8	8
60-69				1		1
50-59			1			
40-49						

3 failed Problem 1 (2 of these passed Problem 2)
5 failed Problem 2 (4 of these passed Problem 1)
1 failed both Problems

TABLE 8

ATC ACADEMY - TERMINAL

INSTRUCTOR RATING

Problem 2

Problem 1	40-49	50-59	60-69	70-79	80-89	90-100
90-100				1	10	11
80-89			2	12	26	8
70-79			3	7	19	5
60-69			1		1	1
50-59					1	
40-49						

4 failed Problem 1 (3 of these passed Problem 2)
6 failed Problem 2 (5 of these passed Problem 1)
1 failed both Problems

TABLE 9

ATC ACADEMY - TERMINAL

PROBLEM AVERAGE (POINTS PLUS RATING)

Problem 2

Problem 1	40-49	50-59	60-69	70-79	80-89	90-100
90-100				2	10	19
80-89			1	2	31	12
70-79			1	5	10	7
60-69			1		2	
50-59						
40-49						

3 failed Problem 1 (2 of these passed Problem 2)
3 failed Problem 2 (2 passed Problem 1)
1 failed both problems
1 failed average of Problems 1 and 2

Automated Assessment of Student Progress and Performance in Radar Air Traffic Controller Training

by James O. Boone, Ph.D.

Introduction

The original simulators used in the Federal Aviation Administration's (FAA) air traffic control training were "patches" developed for the operational automated field systems. The "patches" permitted flexible training at designated positions without interfering significantly with the operational positions. Experiences with these prototype simulators resulted in at least two major notions related to using simulation for radar training. First, the value of computer-driven simulation for training purposes was firmly established. Second, several problems associated with using operational field systems in a training mode were identified. An Institute for Defense Analyses (IDA) study on the training of air traffic controllers discussed some of these problems and suggested that a standardized computer-driven program should be established by the FAA to provide basic radar training. The IDA study further suggested that the radar training should be pass/fail to select out those persons who did not demonstrate the potential to perform proficiently in a radar environment.

In July 1976, engineering requirements were completed by the FAA for a radar training system. During that same month the FAA Administrator approved the procurement and construction of the Radar Training Facility (RTF) to be located at the FAA Academy in Oklahoma City.

In October 1977, the FAA completed a program implementation plan that outlined the development and implementation of the RTF. The contract for the development of the computer-driven simulator training system was awarded to Logicon, Tactical and Training System Division, San Diego, California, in January 1978. Groundbreaking for the construction of the new RTF at the FAA Academy was held on December 22, 1977.

RTF Training System and Laboratory Configuration

The primary objective of the RTF is to closely duplicate the specialized operational environment existing at automated Terminal and En Route facilities as well as have the capability of synthesizing and presenting a wide variety of air traffic control situations. These situations would be based on a reference data base created through scenario programs with a full range of control necessary to establish a realistic simulation of actual aircraft traffic under a variety of conditions.

To accomplish this objective, four independent laboratories are utilized. Figure 1 describes how the laboratories are configured.

Positions. There are Trainee positions and Supervisory and Support positions/stations corresponding to each radar training sector. At a "position," the operating personnel have input/output (I/O) equipment access to the system with associated voice communications. A "station" has no I/O equipment access but is equipped with voice communications for monitoring, instructing, and supervisory functions.

Trainee Position.

1. Radar Control Position (R). The R controller positions (six in each lab) have a display console, (PVD) for En Route and (DEDS) for Terminal. They have associated voice communications. The displays and

1

voice communications are similar to those at field facilities. Displays include maps, weather, aircraft position symbols, alphanumeric readouts, and other digital and symbolic data.

2. Nonradar Controller Position (HO/D). The "D" controller for En Route and the "HO" position for Terminal (six in each lab) have the capability of making and accepting handoffs. This position also permits training for manual or nonradar control by using flight progress strips generated by the flight strip printers.

3. Pilot Position (P). Three pilot positions are associated with each sector (18 in each lab). These positions are in a separate room. Each position operator performs at a console with a tabular display and keyboard for data entry with associated voice communications. These operators simulate aircraft pilots during the exercise by actual responses to ATC clearances/instructions.

4. Ghost Position (G). This position is associated with each R and/or HO/D position. There are six ghost positions in each lab. The position console and display are identical to those of the pilot position. The ghost position operator adds realism to the exercise by performing related functions of adjacent centers, terminals, flight service stations, and positions/sectors. Functions include initiating handoffs, accepting handoffs, and generally ghosting functions of other facilities/sectors.

Supervisory and Support Positions/Stations.

1. Instructor Station (I). An instructor station is provided at each sector (six in each lab). The instructor has voice communication with each student and monitors the overall exercise from behind the trainee positions.

2. Pilot Supervisory Station (PS). This position (one in each pilot room) has voice communications for supervising, monitoring, and instructing operation of pilot positions as well as for coordinating activities with the master instructor station and the system monitor position.

3. Master Instructor Station (MI). This position (one in each lab) controls the exercise within the lab. The position has a tabular display, a data entry keyboard, and associated voice communications with each trainee and with each operator of ghost, instructor, and pilot positions in the lab. The master instructor station will permit setting clock time, starting, monitoring, freezing, backing up, replaying, and restarting the exercise as necessary. The position also provides for data recording and analysis of the exercise.

4. System Monitor Position (SM). One position is provided for each lab. The position will have voice communications with two master instructor positions and two pilot supervisor positions. The position will permit computer operation and operational and maintenance monitoring.

Figure 2 describes the system configuration for operating the positions and stations in each laboratory. The training sectors are controlled by a Digital Equipment Corporation (DEC) PDP 11/60 computer with a PDP 11/34 computer serving as an interface between the PDP 11/60 and the operating positions.

The training process involves three sequential systems of operation: (1) SCENARIO GENERATION ⟶ (2) REAL-TIME ⟶ (3) PERFORMANCE MEASUREMENT. Scenario generation, illustrated in Figure 3, is the non-real-time process of building exercises and evaluation problems for the system. Aircraft characteristics, flight plans, and other essential information of this type are stored in the Universal Data Files (UDF). The exercise is built by first selectively retrieving intermediate files and then creating other intermediate data files from the universal data base through the scenario management program.

1

The real-time component, illustrated in Figure 4, utilizes the scenario management files to generate the actual radar simulation exercise. The real-time component drives the display at the radar position. Aircraft movement is controlled through the pilot and ghost positions according to the instructions the operators of those positions receive from the controller trainee or, in some cases, from a scenario prompt which appears on the cathode-ray tube (CRT) at the pilot or ghost positions. During the operation of the real-time training exercise, all actions taken during the exercise are recorded.

At completion of the exercise, the computer will analyze the recorded actions to determine violations of separation standards and to quantify other pertinent performance information, such as delay times, in order to evaluate the student's ability to move air traffic "safely and expeditiously." The process of student performance measurement is illustrated in Figure 5.

Table 1 contains a list of the measures to be employed in evaluating the students' performance on a given problem.

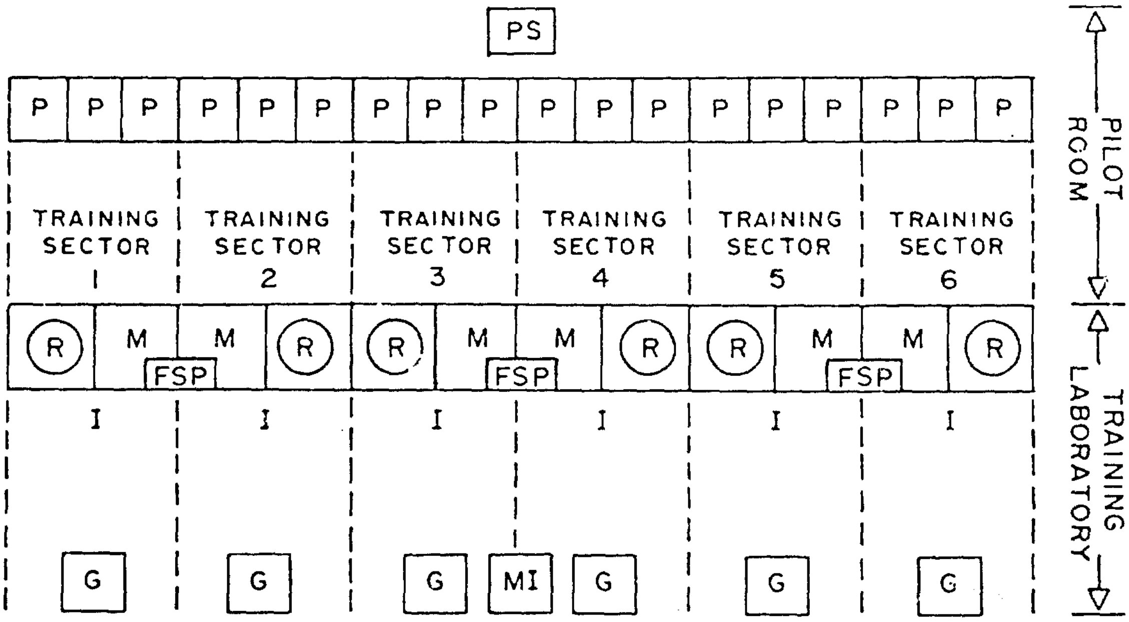
Student Evaluation. The general model for the automated method of evaluation (see Figure 6) is based on the use of latent trait theory applied to adaptive testing in this training situation. It is assumed that each trainee possesses a latent ability in radar air traffic control that is being measured inferentially through testing. This general latent ability consists of several subskills. Assessment of the general latent ability offers an overall evaluation (test score), while assessment of the subskills offers a means to structure a program designed to strengthen a trainee in areas where weaknesses are exposed. In this manner trainees can progress as swiftly as their ability allows, while maximizing their ability through immediate feedback and a tailored curriculum.

Training exercises are executed by means of two files. First, a basic problem scenario file will be built consisting of a series of timed events, such as the entry of various aircraft on specified air routes with a specified flight plan, a Visual Flight Rules pop-up, or an emergency procedure. A second fixed file will contain a list of addressed events and corresponding parameter information that will be used to determine when the event will be introduced into the basic scenario. At specified times during the execution of the basic scenario, an index will be calculated that measures how well the aircraft are being separated and the potential for conflictions. Based on this information and the parameter information from the events in the fixed events file, an event from the fixed events file will be introduced into the problem. This process will be continued until the trainee reaches a plateau or the scenario time limit expires.

During the execution of the problem, cumulative totals will also be calculated on measures such as conflicts, delay time, number of aircraft handled, number and duration of communication transmissions, etc. This information will be stored on the trainee's training record where it can be retrieved immediately in the form of a printout or reviewed later for the purpose of designing the trainee's curriculum. These measures would also be added to a separate master file that contains summary records for all trainees. The master file will be employed to calculate normative information used in comparing a particular trainee's progress at a particular stage in training with that of all others who have been through the program. Figure 6 is a diagram of the evaluation procedure.

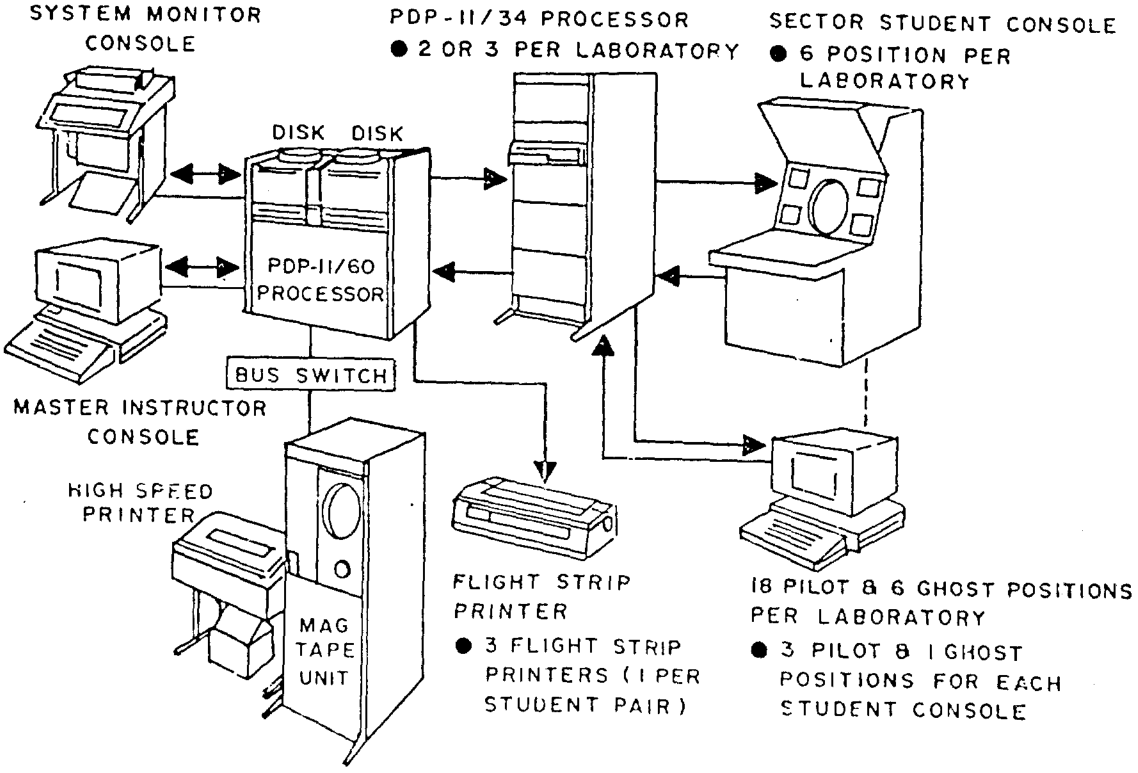
Conclusions

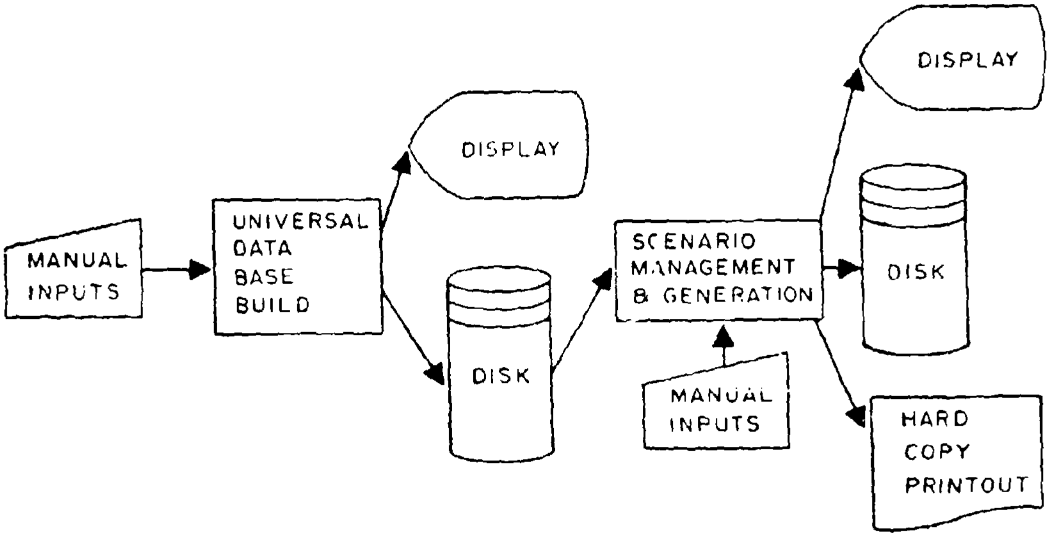
The philosophy behind the RTF is to place the ATCS trainee in the automated environment for a brief and intense period of training and to rigorously evaluate how well the trainee operates. If the workload or complexity of the tasks is beyond or outside of the trainee's aptitudinal capabilities, it is directly observed. The observation is systematic and contains sequential steps. The trainee is given direct feedback at each step in training. If a trainee fails to proceed at a successful rate, the trainee may be screened from the program. Early detection of those who are unable to operate successfully in the automated radar ATC environment accomplishes at least two things. First, it allows the failing trainee to enter another occupational field much sooner, and second, it lessens the impact of automation on ATCS personnel by helping to insure that the persons operating in the automated radar environment are better matched with the job requirements. It is believed that employing this facility as a mini-laboratory for observing ATCS trainee behavior can serve as a major impetus in lessening the impact of automation on ATC personnel.

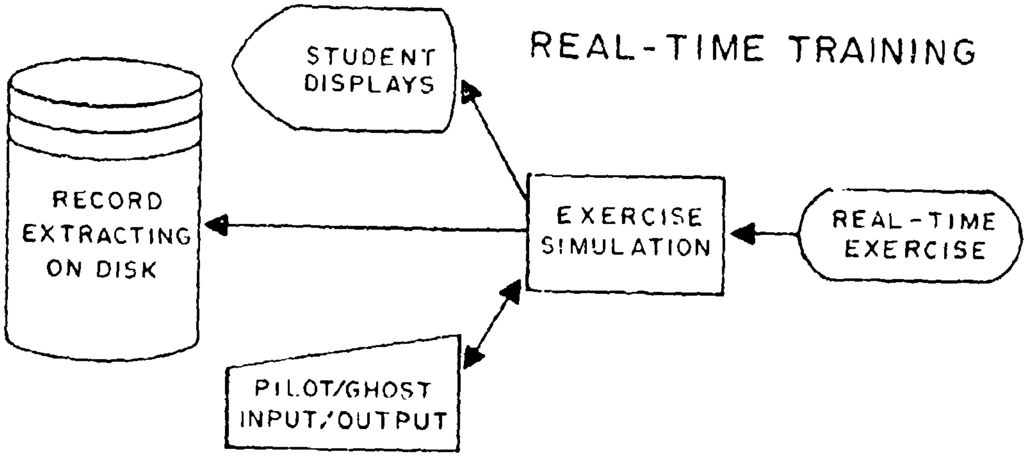


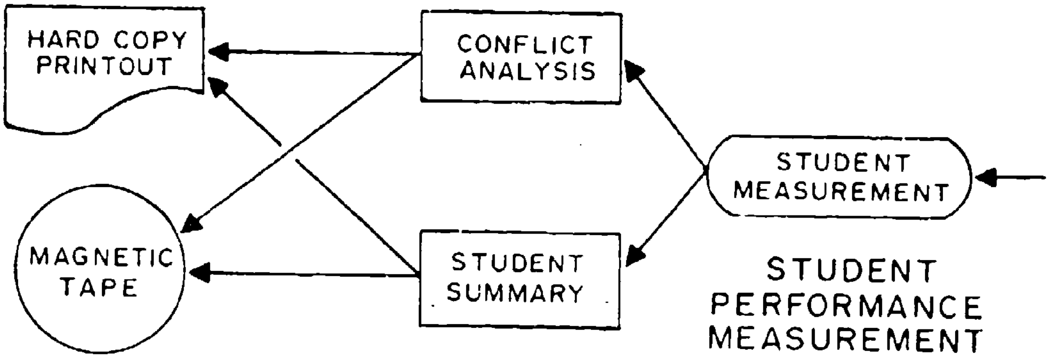
P - Pilot Console
G - Ghost Console
R - Radar Training Console
M - Manual Controller Trainee Position
En Route "D" Controller or
Terminal "HO" Controller

MI - Master Instructor Position
PS - Pilot Supervisor Station
I - Instructor Station
FSP - Flight Strip Printer


SYSTEM MONITOR CONSOLE
PDP-11/34 PROCESSOR
● 2 OR 3 PER LABORATORY
SECTOR STUDENT CONSOLE
● 6 POSITION PER LABORATORY
DISK DISK
PDP-11/60 PROCESSOR
BUS SWITCH
MASTER INSTRUCTOR CONSOLE
HIGH SPEED PRINTER
MAG TAPE UNIT
FLIGHT STRIP PRINTER
● 3 FLIGHT STRIP PRINTERS (1 PER STUDENT PAIR)
18 PILOT & 6 GHOST POSITIONS PER LABORATORY
● 3 PILOT & 1 GHOST POSITIONS FOR EACH STUDENT CONSOLE


MANUAL INPUTS
UNIVERSAL DATA BASE BUILD
DISPLAY
DISK
SCENARIO MANAGEMENT & GENERATION
MANUAL INPUTS
DISPLAY
DISK
HARD COPY PRINTOUT


REAL-TIME TRAINING
STUDENT DISPLAYS
RECORD EXTRACTING ON DISK
EXERCISE SIMULATION
REAL-TIME EXERCISE
PILOT/GHOST INPUT/OUTPUT



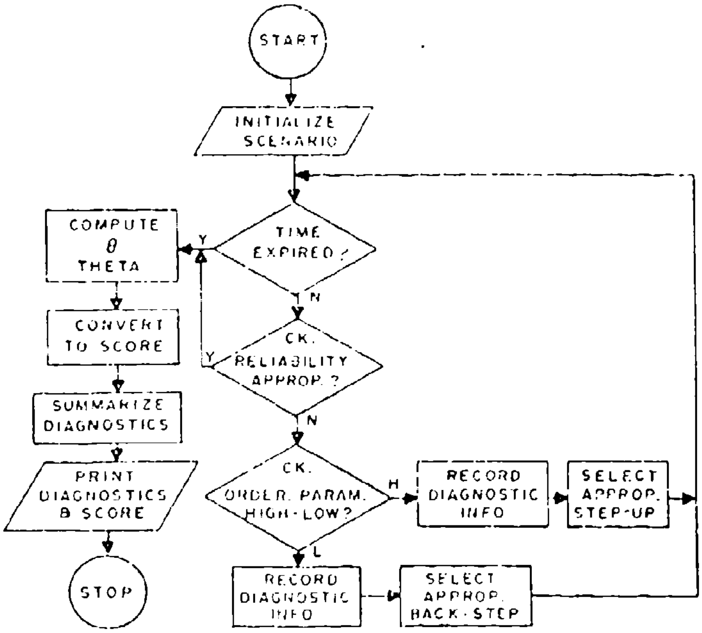


Figure 6. Diagram of the evaluation process.

Table 1. List of RTF Measures

1. Number of aircraft in the sample.
2. Ideal aircraft time in system (based on filed flight plan).
3. Ratio of the ideal aircraft time in system and the number of aircraft in the sample.
4. Number of completable flights.
5. Data period duration.
6. Number of arrivals.
7. Number of departures.
8. Arrival/departure ratio.
9. Arrival rate scheduled per hour and departure rate scheduled per hour.
10. Conflicts--terminal (3 nautical miles (NMI)).
11. Conflicts--en route (5 NMI).
12. Number of delays (start time).
13. Delay time (start time).
14. Number of delays (hold and turn).
15. Delay time (hold and turn).
16. Number of delays (arrival).
17. Delay time (arrival).
18. Number of delays (departure).
19. Delay time (departure).
20. Number of delays (total).
21. Delay time (total).
22. Aircraft time-in-system (real).
23. Number of aircraft handled.
24. Number of completed flights (total).
25. Number of arrivals achieved.
26. Arrival rate achieved per hour.
27. Number of departures achieved.
28. Departure rate achieved per hour.
29. Number of air-ground contacts.
30. Air-ground communications time.
31. Number of altitude changes.
32. Number of headings changes.
33. Number of speed changes.
34. Number of path changes (altitude, heading, and speed).
35. Number of handoffs.

Development of the Multiplex Controller Aptitude Test

by Evan W. Pickrel, Ph.D., and John T. Dailey, Ph.D.

Introduction

The Multiplex Controller Aptitude Test is a new measure created for the screening of air traffic controller applicants. It presents job sample items from the controller activity, shows continuing air traffic movement across a simulated radar scope and asks questions about that traffic. Traffic changes are sequential, each item using its own illustration and flight data table to describe the progression of that traffic across the scope. About 40 percent of the 55 questions were created to measure the prediction of traffic conflicts, violations of separation standards. The others were initially designed to measure the traditional types of aptitudes such as table reading and interpretation of data, spatial visualization and orientation, relative target movement and arithmetic reasoning regarding separation times and distances. These latter questions have the characteristics of two-dimensional separation items. The number of items to be included in each aptitude area was estimated during early studies relating item types to total test homogeneity.

The format for item sequencing is a departure from the usual aptitude battery design practice of clustering items into homogeneous subgroups. In this test, the items alternate from one type to another and spiral to increasing levels of difficulty, a mode of presentation found only in a few tests such as the Stanford-Binet. A result shown statistically is that nonconfliction items included for measurement of aptitudes represented in today's selection battery show unexpectedly high commonality with the new air traffic items requiring detection of impending conflictions. The resulting measure shows promise of consistently producing higher validity coefficients than possible with the current test battery.

Background

The current selection test battery for screening air traffic controller applicants has been operational since 1964. In 1970, the Federal Aviation Administration contracted with specialists in the field of personnel selection to search for ways to alleviate problems being experienced in selection and retention of air traffic controllers. They identified other available tests that might increase the predictive validity of this battery, plus areas for the construction of new tests. One of these (Buckley, Note 1) utilizes motion picture films to present simulated air traffic in real time as it crosses an actual controller display. An observer is supposed to predict violations of separation standards, or traffic conflicts, as early and accurately as possible. This instrument added significantly to a composite for predicting on-the-job success (Milne, Note 2), but the scoring system is cumbersome, requiring an extensive computer program. In a rescoring and analysis of the 1971 data, it was found that most of what can be measured in each 45 minutes film could be derived from only seven or eight pairs of true conflictions. A simplified scoring system was developed that could be done easily by hand by untrained clerks, but the equipment and space required for motion pictures tests generally is not available in Civil Service Commission testing situations. Therefore, test development procedures were initiated to derive measures of this same skill in a format that meets operational needs.

Test Development

Experimental administration of three available motion picture forms of CODE were carried out to learn something about the test and search for an acceptable format for operational use. The motion picture forms provide an unstructured, free response mode in which the observer predicts and records potential conflicts as early and accurately as possible. The observer has much idle time during testing, as traffic flow is in real time, the traffic presented is light, and aircraft widely separated. In this sense, the setting resembles a common air traffic situation, but an extended amount of testing time is required to present only a small number of confliction items. Better use of examinees' time would be desirable.

From early analyses, it was found that identification of potential conflicts was easier, quicker, and more accurate after observers adjusted to the scope and its targets, when only a few targets were in near-confliction, and when their rate of closure was slow. Items were more difficult when the observers were first exposed to the scope and its targets, when multiple targets were in near confliction and when the rate of closure among targets was rapid. False positive conflict items did not discriminate and generally, for true conflicts, the greater the lead time available, the greater an item's discriminating power. Journeyman controllers identified potential conflicts sooner than developmental trainees, and identified almost all potentials that became real conflicts. Developmental controllers generally missed a number of real conflicts, were slower than journeyman in calling them out, and frequently, did not attend to aircraft altitude separations. However, the range in test performance among developmentals was great, and some performed as well as journeymen. Some confliction items in these films were too easy and some had a negative relation to a developmental-journeyman criterion dichotomy.

An initial change during test development was to structure the items, using slides to present questions on the screen above the film. Mean response times for developmental and journeyman controllers to react to each conflict in the free response versions were determined. Journeyman mean times were selected to be determiners of aircraft positions when presenting confliction questions in the structured versions, for this maximized the discriminating power of each item. Two-choice conflict items, asking whether a pair of aircraft would conflict, were presented for 30 seconds when unexpected changes occurred among targets, such as when new aircraft entered the picture. Four-choice conflict items, asking which pair would conflict if a confliction occurs, were presented for 45 seconds when traffic changes were slow. A "None of these" response was introduced to permit inclusion of nonconflictions that might be predicted to be conflictions, and this was used as the fourth alternative in all four-choice confliction items. Items were assembled into the described format as an experimental film/slide test format but they didn't utilize half the available testing time. The large amount of idle time, between questions regarding conflicts, provided much opportunity for observers to change their answers to earlier questions as the aircraft approached each other and the correct answers became more obvious. In the free response film version, this problem is controlled by a requirement for entries from a coded clock onto the computer-scored answer sheet whenever potential conflicts are reported. It was found with the film/slide version, that presentation of a new item every 45 seconds keeps examinees so busy that they have no opportunity to make such changes. However, more test items are needed if this type of control is to be used.

CODE films provide pictures of simulated traffic moving across a radar scope plus a table which includes detail data on each aircraft, identifying the target, its altitude, speed, and route on the scope. The scope uses lines to represent airways or highways in the sky, with alphabetical identifiers of starting, ending, and intersecting points on the airways. The top of the scope is North, and a mileage scale is provided at the bottom. Ample information is presented to prepare a variety of questions related to a controller's activity. It appeared possible that most factors used in the current controller selection battery could be measured within this simulated air traffic setting.

Items were written that utilized the available detailed information to measure such aptitudes as direction following, table reading, interpretation of data, spatial visualization and orientation, estimation of distances and relative target movements, and arithmetic reasoning regarding separation distances and times. Some items included were of a very simple nature, and others were written in a multi-factor format to increase their level of difficulty. For example, the directions for this type of test require instruction on how to read the table, so initial table reading questions were included that were of the very easy, instructional type. Awareness of distances across the scope, reading the table to determine the speeds of aircraft, and mathematical computation to determine their rates of closure (all of these being related to horizontal separation) were required to solve a complex problem such as estimating the travel time in minutes between two aircraft at a given moment.

The relation between item types and total test homogeneity was determined, and this ratio was used to determine the number of items per type to include in the test. A result, for example, was inclusion of twice as many target time-distance separation items as compass heading items in the test. The order of placement of conflict items had already been established by using the mean response time of journeyman controllers when targets were at certain locations. New aptitude test items were placed in the remaining positions, alternating from one type to another and spiraling to increasing levels of difficulty as testing progressed. About 40 percent of the items were included to measure the prediction of conflicts, and the other 60 percent to measure the various aptitudes. As CODE films 4, 6, and 7, each provided a different pattern of air traffic, three alternate forms of the test, MCAT 4, 6, and 7 were prepared.

Table 1 describes a variety of the forms given to this job sample test as it evolved from an early, unstructured free response motion picture test for prediction of traffic conflictions. In its current structured, multiple choice, paper and pencil format, MCAT includes a wide variety of questions about the scope, traffic characteristics and separation, as well as traffic conflictions.

Versions of the CODE test, the film/slide version of MCAT, and other selection measures were administered to students at the U.S. Navy Air Traffic Controller Training School, Memphis during the week of August 18-22, 1975. Grades of these students were obtained as they progressed through classroom and laboratory training and passed or failed the course. Distribution statistics, intercorrelations and rotated factor loadings of the various tests with each other and with such school grades as laboratory flight plans, control tower, and radar control problems were obtained, as those grades provide the greatest range in criterion scores for differentiating performance among FAA student controllers. The 30 minute film/slide version of MCAT generally had correlations with criterion variables that are as high or higher than those between the Free Response CODE test and those same criterion variables. These results encouraged further development of the Multiplex Test.

A next step in test development was to convert the slides plus film presenting moving targets to a "slide only" presentation, in essence capturing pictures of the scope with targets in the same position as when each question appeared on the screen. Pacing of slide presentation was the same as in the film plus slide version, so the amount of target movement from question to question was unchanged. This format was easier for test administrators, since they no longer had to cope with film projection problems, and the stationary target presented by slide seemed easier to read than the moving targets presented by film. Persons taking the experimental version in this slide format did grumble that the test seemed to take control over their time, as each item had to be completed quickly before it left the screen and a new item appeared. This task characteristic has commonality with controller situations in the real world, as traffic movement tends to control their involvement and pace of work. Another merit of this version would be its ease for transition into a paper and pencil format that would meet Civil Service Commission requirements for use in their very decentralized field testing situations. Flim plus slide and "slide only" versions of the Multiplex Controller Aptitude Test plus other selection measures were administered to students at the USAF Air Traffic Controller Technical School, Keesler Air Force Base, Mississippi, during the week of October 20-24, 1975. Grades of these students were obtained as they progressed though classroom and laboratory training and passed or failed the course. Distribution statistics and intercorrelations of the various tests with school grades were determined. The slide version of MCAT generally showed higher correlations with the criteria than either CODE 7 (Free Response) or the combination film/slide version.

A next step in test development was to print the slide versions in paper and pencil format. The slide version and combination film/slide version permitted time control at the individual item level, allowing 30 seconds for response to two-choice items and 45 seconds for response to four-choice items. A segmented paper-and-pencil version was prepared which extended time controls beyond the item level, providing 5 minutes to work on each cluster of items. Five clusters were formed for MCAT 606B, and six for 706B. A cluster could hold as many as 11 items when they were all of the two-choice type. A further extension of time controls was a paper-and-pencil version that allows uninterrupted work throughout testing time, with announcements when 15 minutes have elapsed and when only 5 minutes remain to work the test. Three parallel forms of the written test were prepared, MCAT 406 including 41 items, 606 containing 43 items, and 706 containing 53 items. Time limits are 25 minutes plus directions for MCAT 406 and 606 and 30 minutes plus directions for MCAT 706.

These forms were available by January 1976 and administered to several populations, including; (a) entering students at the ATC Academy during all of that year, (b) about 7000 ATCS applicants to the Civil Service Commission from the Southern and Eastern Regions during the Fall of 1976 and Spring of 1977, and (c) all active and inactive air traffic controller specialists included in an extensive contractural study to determine the potential of an experimental test battery and other measures to predict the success of applicants for air traffic controller work (Mies, Note 3). Throughout the reports on these studies, the MCAT appeared to be the most promising, of all tests included, for initial screening of ATCS applicants.

Each form then was lengthened to 55 items. MCAT 407 includes 23 conflict and 32 aptitude items, MCAT 607 contains 22 conflict and 33 aptitude items, and MCAT 707 contains 23 conflicts and 32 aptitude items. Time limits are 35 minutes for each of these tests. These forms were available by January 1977 and administered to several populations, including; (a) entering students at the ATC Academy during all of that year and until June 1978, (b) students at various colleges and universities, and (c) students at the U.S. Army Air Traffic Controller School, Ft. Rucker, Alabama. Each of the Army ATC students were administered all three forms of the test, and differences in mean performances among the forms were found that could effect form comparabilities. It was also found that student performances improved considerably on the second form they were administered, indicating that learning was still taking place, but there was very little increase in performance on the third form. Relations between ATC Academy student performances and Pass-Fail continued to show MCAT as the most promising available measure for screening air traffic controller applicants.

In discussions with Civil Service Commission personnel, it was found that they would require many alternate forms of the test before placing it in operational use, to meet such requirements as retest of persons requesting it, and variation of test content for control over compromise. Therefore, each of the available forms was split into two parts by placing the odd items in one part and the even items into a second part. Some changes in item sequence proved necessary to equate the parts for such things as number of conflictions. Then new two-part forms were created by combining the odd-item form from 407 with the even-item form from 607, the odd-item form from 607 with the even-item form from 707, and so on. Each part now required a person to learn its unique pattern of air traffic and make predictions as to its behavior and potential conflictions. The six available parallel forms have been administered to students entering the ATC Academy since June 1978, with each student taking two forms of the test plus an ATC Occupational Knowledge Test. Many scoring procedures have been tested for both the reliability of the derived scores and for their relation to Pass-Fail performance in the ATC Academy. These are described in the section that follows.

Scoring Method

Many different scoring methods have been explored during development of the test. The test could be scored for Rights and for Wrongs, with separate scores for the different types of items, the Conflicts and the Aptitudes, and with differential combining weights for these various types. The test could be scored for the rate of improvement from early parts to later parts of the test, with differential combining weights for the different parts. Early item analyses revealed that the various types of conflict and aptitude items were statistically quite homogeneous, but for some time separate scores were maintained for Conflicts and Aptitudes. Various multiple regression studies indicated that little was to be gained from this, so performance on all types of items were combined in a single scoring.

Table 2 presents the intercorrelations for Rights and Wrongs raw scores on two forms of MCAT and their correlations with course Pass-Fail for 617 incoming students tested at the ATC Academy from June through November 1978. The correlation for the sum of the Rights scores for the first and second administration is .55, and the multiple correlation of optimally weighted Rights and Wrongs scores is .557. A study of the Rights and Wrongs score intercorrelations and validities for Pass-Fail shows that very little is to be gained through use of a combination of Rights and Wrongs scores instead of a Rights only. Thus, it is recommended that the MCAT be scored Rights Only or R-W/3 if thought desirable for administrative reasons.

Table 3 presents intercorrelations of rights raw scores for each half or part-test of two forms of MCAT and their biserial correlations with course Pass-Fail for 617 incoming students tested at the ATC Academy from June through November 1978. The part score intercorrelations are relatively low, indicating that the full length testing time of 90 minutes is essential to obtain an adequate level of reliability. The biserial validities for the four part scores are .35, .45, .48 and .43; the validity for parts 1 & 2 and parts 3 & 4 is .54; and the validity for the sum of the four parts is .55. The multiple correlation of optimally weighted Parts 1 & 2 and Parts 3 & 4 is .566, an extremely small increase over the .55 validity for the sum of the four parts. The 45 minutes first form test validity is .45; the 90 minutes test's validity is .55. It is probably best to weight each part equally and use only total Rights score for the entire 90 minutes testing of MCAT.

Reliability

The development of various forms of MCAT has resulted in the reporting of reliability coefficients that vary with the restriction in range of the scores for the populations reported. Table 3 presents a total Rights raw score test-retest correlation of .60 between comparable 45 minutes forms of MCAT when administered to entering ATC students. A doubling of test length, providing a full length testing time of 90 minutes, may be expected to increase the reliability coefficient to .75 for this type population, and provide an adequate level of reliability for use in the operational situation.

Validity

During evolvement of the various forms of MCAT, test validation studies have used the criterion of "success" in air traffic control work. Success as defined here for ATC applicants is hierarchical, including (1) satisfactory completion of the initial, formal training program (2) satisfactory performance on the job, an (3) progression or upward mobility within the ATC system. Another, element, (4) attrition, may be a measure of nonsuccess. Those enrolled in the initial, formal training program are a highly select group, for they scored high enough on the initial selection tests to be hired, but group membership becomes even more selective as those who fail to learn or perform adequately are separated. This selection process continues as they progress up the ATCS career ladder and are evaluated for satisfactory performance on the job and advancement into the more demanding and higher paid positions within the ATC system. The range in scores on these selection tests will be greatest for the group when they first enter formal training, but become ever more restrictive with career progression. As this restriction in range may have a direct effect on the size of the validity coefficients, the validities should be highest when the group is in initial training and become lower for that portion of the group which progresses up the career ladder.

The MCAT scores have constantly produced higher validities for success in training than other candidate tests for the OPM selection battery. MCAT has been the prominent predictor in most regression equations for prediction of job performance success and ATC progression. (Mies, Note 3). An aggregate criterion of ATC "success" has been constructed from combinations of the four individual criteria (training, on-the-job performance, progressions, and attrition) and provided a five point scale value for ATC success. The MCAT again has been found to be the major factor in the validities derived from the multiple regression analysis.

Table 3 presents validities for MCAT when two forms were administered to 617 incoming student at the ATC Academy from June through November 1978. Validity for the total Rights score for the first form is .45, for the second form is .54, and the validity for the sum is .55. The multiple correlation of optimally weighted parts is .566. As these data have been gathered and related to performance at the academy it has been shown that a sum of scores earned on two forms of the test has an unusually high relation to pass-fail in the course, and a significantly higher relation than scores on a single form. In study of individual's performances on parts 1 through 4 of the two forms taken, it was found that performance gains were greatest with the first three parts, but were reaching a plateau with part 4. This and other analyses indicate that each part seems to be a distinct learning problem, involving such activities as learning the distances to various points on the scope, patterns of the various routes, the detail flight characteristics of each unique sample of air traffic to be controlled, predicting two dimensional conflicts when answering questions such as time and distance between aircraft, and predicting three dimensional conflicts. As air traffic is unique to each part, it is felt that this measurement of a person's ability to learn new and unique patterns of air traffic, and make predictions as to their behavior and potential conflictions, is the heart of the successful prediction story being described. In the data presented here, the test validity for administration of the first 45 minutes form is .45, and test validity for the 90 minute administration of two forms is .55. It is recommended that a Total Rights score for a 90 minute administration of the test be used in the operational situation.

Summary

A new test has been developed for initial screening of FAA Air Traffic Controller applicants. Its content includes the traditional types of aptitude test items found in today's initial screening battery. In addition, it includes a measure of the ability to identify potential conflicts in air traffic, a skill that has been demonstrated experimentally to have significant relation to success in the FAA ATC specialty. All test questions are presented in an air traffic control setting, which gives them a job-related appearance not found in today's selection battery. Alternate forms of the test have been developed in a paper-and-pencil format to meet OPM needs in their decnetralized testing program. The test has been administered experimentally to groups whose abilities approximate those of the applicant population, and correlations during a full length testing time of 90 minutes indicate that it has satisfactory reliability characteristics. Selection using scores from this test meets the OPM-recommended clearly model for test fairness. It has been administered experimentally to students entering the new FAA ATC Academy and personnel on the job at operational facilities, and constantly produced higher correlations with ATC success than any other tests used in the validation studies. The available data indicate that this new and customized instrument, when used in combination with other selected measures, promises to be a significant improvement over the existing battery for screening FAA Air Traffic Controller applicants.

TABLE 1

EVOLUTION OF MCAT

A job sample test presenting air traffic movement across a simulated radar scope and asking questions about that traffic.

UNSTRUCTURED FREE RESPONSE, predicting traffic conflicts as early and accurately as possible.

Motion Picture, CODE 4, 6, and 7. Scoring: 1971-40 items, 1974-Stanine, 1975-corrected item scores- best items.

STRUCTURED MULTIPLE CHOICE, answering conflict, separation and traffic related questions as quickly and accurately as possible. Approximately 40% are conflict items.

Film Slide, time controlled for working each individual item. July 1975
MCAT 406 FS- 41 items, 606 FS- 43 items, 706 FS- 53 items.

All slide, time controlled for working each individual item. October 1975
MCAT 406 AS- 41 items, 606 AS- 43 items, 706 AS- 53 items.

Paper and Pencil, time controlled for working total test. January 1976
MCAT 406A- 41 items, 606A- 43 items: 25 min. testing time.
MCAT 706A- 53 items: 30 minutes testing time.

Paper and Pencil, time controlled for working each five minute group of items.
MCAT 606B- 43 items, 706B- 53 items. April 1976

Paper and Pencil, time controlled for working total test- 35 minutes.
MCAT 407A, 607A, 707A: each 55 items. January 1977

Paper and Pencil Two-parts, each with a unique pattern of air traffic.
Timing: Part 1-20 minutes, Part 2-15 minutes. June 1978
MCAT 4o6e, 4e7o, 6e7o, 6o7e, 7e4o, 7o4e: each 55 items.

TABLE 2

INTERCORRELATIONS AND VALIDITY COEFFICIENTS
MCAT* RIGHTS AND WRONGS, ATC ACADEMY PASS-FAIL
617 STUDENTS ATC ACADEMY
JUNE - NOVEMBER 1978
(Decimal points ommitted)

	1	2	3	4	5	6	7
First MCAT Administered							
1. Rights		-24	60	-21	91	-27	45
2. Wrongs			-24	48	-27	85	-22
Second MCAT Administered							
3. Rights				-59	87	-49	54
4. Wrongs					-43	87	-32
Sum First & Second MCAT							
5. Rights						-41	55
6. Wrongs							-32
ATC Academy							
7. Pass-Fail							
Multiple Regression							
Rights & Wrongs							557

*Forms 4o6e, 4e6o, 6o7e, 6e7o, 7o4e, 7e4o raw scores directly combined, all forms

TABLE 3

INTERCORRELATIONS AND VALIDITY COEFFICIENTS
MCAT* PART TEST SCORES, ATC ACADEMY PASS-FAIL
617 STUDENTS, ATC ACADEMY
JUNE- NOVEMBER 1978
(Decimal points omitted)

	1	2	3	4	5	6	7	Course Pass-Fail
First MCAT Administered								
1. First Half, Rights		58	89	48	42	51	80	35
2. Second Half, Rights			89	48	49	56	83	45
3. Total Rights				89	89	60	91	45
Second MCAT Administered								
4. First Half, Rights					52	85	76	48
5. Second Half, Rights						89	77	43
6. Total Rights							87	54
First & Second MCAT								
7. Total Rights								55
Multiple Regression								
First & Second MCAT Rights								566

*Forms 4o6e, 4e6o, 6o7e, 6e7o, 7o4e, 7e4o raw scores directly combined, all forms

COMPARISON OF NUMBER OF SMS INCORRECT ITEM RESPONSES WITH TEST ITEM STATISTICS[1]

Conrad G. Bills
and
Gary D. Stanfill

USAF Occupational Measurement Center
Randolph Air Force Base, Texas

A paper presented at the 21th Annual Conference of the
Military Test Association
October 1979

[1] The views expressed in this paper represent those of the authors and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

COMPARISON OF NUMBER OF SMS INCORRECT ITEM RESPONSES WITH TEST ITEM STATISTICS

Conrad G. Bills, Capt, USAF and Gary D. Stanfill, Amn, USAF

USAF Occupational Measurement Center
Randolph AFB, TX 78148

Specialty Knowledge Tests (SKTs) written at the USAF Occupational Measurement Center are used by the Air Force in the Weighted Airman Promotion System (WAPS). Subject-matter specialists (SMSs) are sent by temporary duty orders to the Center to write the SKTs for their career field. Generally, four SMSs write two SKTs, one for promotion testing to E-5 and the other for promotion testing to E-6/7. The SMSs write the test questions. These test questions are evaluated by test psychologists according to standard test writing criteria. Once a test psychologist approves a test item, the item is evaluated by a more experienced review psychologist who assures the quality of the item. The SMSs select from the item pool the questions for each 100-item SKT. An SKT undergoing major revision is composed of 50 or more new items. These SKTs are critically reviewed by the SMSs, the test psychologist, the review psychologist, and also by the senior review psychologist before publication. SKTs written during one calendar year are generally administered during the following year's promotion testing cycles. Personnel competing for promotion are administered the SKT for their career field. The test scores achieved by these personnel are included in the WAPS formula with other pertinent information, e.g., supervisor evaluation and time in grade. WAPS scores are computed and rank ordered, then a cutoff score is determined based on the promotion quotas. Personnel with WAPS scores above the cutoff score are promoted to the next higher grade. For personnel competing for promotion, their SKT score is the most important part of WAPS because they can have an impact on the overall WAPS score by individual study.

Since SKT scores are a part of the WAPS score, this prohibits pretesting before the promotion test administration. Pretesting would give a decided advantage to the portion of the test population used for validation. Additionally, the importance of keeping current the specialty knowledges covered by the tests dictates that the time between test construction and administration must be kept to a minimum, precluding time for pretesting. Therefore, the measure of item quality for new test questions is the subjective evaluation of the SMSs and the respective test psychologists. The complete item statistics for new test questions are not available until after all promotion scores for the testing cycle have been computed, a period of one to two years from the time the new test questions are written. Since pretesting is not considered feasible, there is a method for quasi pretesting during the test construction project that could be considered useful for item validation.

Once SKTs are prepared as 100-item tests, these tests are administered to the SMSs that wrote them. Since SMSs do make incorrect responses in varying degrees, from one to all four SMSs missing a given item, it is possible to quantify test item quality. The purpose of this study was to compare the number of SMS incorrect responses for each item to the subsequent item difficulty and discrimination indexes. This paper is a progress report on the first comparison since the item statistics became available.

METHOD

During the calendar year 1978 there were 108 tests which underwent major revision or were written for the first time. Of these 108 tests, 45 SKTs were selected from four different test construction periods, one in each quarter of the year. SKTs were selected to be representative of the 108 tests. Each team of four SMSs wrote an E-5 SKT and an E-6/7 SKT. After each SMS team was administered the test they had written, the number of SMS incorrect responses, i.e., one through four, was compiled for each test question.

For this study, test item statistics were available for 25 of the 45 tests selected (23% of the 108 tests in 1978). Eleven were E-5 SKTs, for a total of 1076 items, and 14 were E-6/7 SKTs, a total of 1379 items. (Item statistics were not available for 45 items-- zero to five items from each test). Career fields represented were as follows: Motion Picture, Airfield Management, Telecommunications, Radio Relay Equipment, Missile System Maintenance, Avionic Sensor Systems, Telephone Switching Equipment, Jet Engine Mechanic, Construction Equipment, Management Analysis, Personnel, Medical Administration, and Preventive Dentistry.

The comparison of number of SMS incorrect responses was made with the corresponding item difficulty and discrimination indexes. The number of SMS misses was combined for one and two (1/2) incorrect responses and also for three and four (3/4). Analysis included the F-test for homogeneity of variance, Chi-square and the coefficient of colligation (Q) for consistency of data, and Jenkins index of covariation (JIC) for magnitude of data (Jenkins & Hatcher, 1976).

RESULTS

The variance comparison of the 1/2 incorrect response group with the 3/4 group indicated that the variances were heterogenous for both E-5 SKTs (difficulty index, $F = 1.2$, $p < .05$; discrimination index, $F = 2.6$, $p < .05$) and E-6/7 SKTs (difficulty index, $F = 1.1$, $p < .05$; discrimination index, $F = 2.1$, $p < .05$).

Except for the E-5 SKT discrimination index, the Chi square indication of consistency between the 1/2 incorrect response group and the 3/4 group was significant (Table 1). Each of the E-6/7 SKT comparisons using Chi square were significant (Table 2).

Table 1

Comparison of Number of SMS Incorrect Responses with Item Difficulty and Discrimination Indexes for E-5 SKTs Using Chi Square

Index		Incorrect Responses			
		1/2	3/4	X^2	p
Difficulty	Mean	42.73	32.56	8.52	.01
	Range	2 to 95	2 to 74		
		(93)	(72)		
Discrimination	Mean	13.93	8.35	2.07	NS
	Range	-28 to 51	-10 to 24		
		(79)	(34)		
	0 or neg	10.3%	26.5%		
N		400	34		
% of Total Items		37.1%	3.2%		

Table 2

Comparison of Number of SMS Incorrect Responses with Item Difficulty and Discrimination Indexes for E-6/7 SKTs Using Chi Square

Inuex		Incorrect Responses			
		1/2	3/4	X^2	p
Difficulty	Mean	47.57	35.04	8.54	.01
	Range	5 to 95	1 to 69		
		(90)	(68)		
Discrimination	Mean	18.26	13.60	8.27	.01
	Range	-31 to 60	-9 to 40		
		(91)	(49)		
	% 0 or neg	9.6%	13.2%		
N		594	53		
% of Total Items		43.1%	3.8%		

The Q (difference of cross products divided by sum of cross products) indications of data consistency between the 1/2 and the 3/4 incorrect response groups were significant for both E-5 and E-6/7 SKTs (Table 3).

Table 3

Comparison of Number of SMS Incorrect Responses with Item Difficulty and Discrimination Indexes for E-5 and E-6/7 SKTs Using Coefficient of Colligation (Q) and Jenkins Index of Covariation (JIC)

Level	Index	Q	p*	JIC	p*
E-5	Difficulty	.483	.005	.623	.005
	Discrimination	.250	.01	.350	.005
E-6/7	Difficulty	.396	.005	.590	.005
	Discrimination	.400	.005	.505	.005

*Fisher's Table of Correlation Coefficients

The JIC (range in means divided by range in N) indications of data magnitude between the 1/2 and the 3/4 incorrect response groups were also significent for both E-5 and E-6/7 SKTs (Table 3).

DISCUSSION

The number of SMS incorrect responses to the items of the SKT they have written was found to be a quasi pretest that could be used for item validation. The covariation of number of SMS incorrect responses with the item difficulty and discrimination indexes indicated that as the number of SMS incorrect responses increased, the item difficulty index increased and the discrimination index decreased. The more SMSs that missed a test question, the more difficult the item and the more unlikely the item will discriminate between the upper and lower portions of the test examinee population. This position was strengthened by the magnitude of the difference between the 1/2 and the 3/4 incorrect response groups. This demonstrated that the 3/4 incorrect response group was the strongest indicator of poor quality test questions.

Since the 3/4 incorrect response group is the strongest indicator of poor test item quality, then test items that are missed by three or more SMSs should be reevaluated.

However, a decision to automatically replace the 3/4 group items is not fully justified. There were items in the 3/4 incorrect response group that were well within the acceptable limits for both difficulty and discrimination indexes. The chance for error in predicting the item quality is high because the quasi pretest population is limited to four SMSs. Also, the SMSs are E-7 selectees or higher in grade and therefore do not represent the test examinee population. Yet, even with these limitations, this quasi pretest is better than subjective judgment. Therefore, the procedure for collecting number of SMS missed data should be instituted as a requirement for each major test construction project.

After the administration of the SKTs to the SMSs, the test psychologist should tally the number of incorrect responses for each test question. The test psychologist and the SMSs should reevaluate each item with three or more incorrect responses. These items should be simplified, corrected, or replaced except when there is strong justification to let the item remain unchanged. Since this study was the initial evaluation, there should also be concurrent reevaluation of the procedure.

When the circumstances dictate that traditional methods of pretesting are prohibited, the evaluation of items is left to subjective judgment. In this context the use of this quasi pretest is an acceptable procedure for test item validation.

REFERENCES

Jenkins, W.O. and Hatcher, N.C. The Design of Behavioral Experiments. Auburn University at Montgomery, AL (unpublished manuscript), 1976, 255-270.

CRITERION-REFERENCING THE APPRENTICE KNOWLEDGE TEST TO AIR FORCE TECHNICAL SCHOOL GRADUATES*

Captain Michael D. McMillan

and

Major Martin J. Dittmar

USAF Occupational Measurement Center
Randolph Air Force Base, Texas

ABSTRACT

The Apprentice Knowledge Test (AKT) is a 100-item multiple-choce test used as a relative measure of specialty knowledge at the semiskilled level. The AKT is used primarily to (a) serve as one index of specialty knowledge attainment in OJT and (b) identify the relative knowledge of enlistees as an aid in the determination of suitability to bypass technical training. Following detailed research, the decision was made to criterion-reference the AKT to the performance of AF technical school graduates. Preliminary test results are consistent with the research. This paper is a follow-on to one presented at last year's MTA conference and discusses procedures developed to administer the criterion-referencing program, the problems which arose in implementing the program, and predicted benefits of the program.

*This paper was presented but, due to its unavailability at the time of printing, only the abstract is reproduced here.

THE USE OF COMPRESSED SPEECH IN SELECTING MORSE CODE OPERATORS

David E. Servinsky
Department of Defense
Fort George G. Meade, Maryland 20755

INTRODUCTION

The ability to learn International Morse Code (IMC) is apparently a special aptitudc unrelated to other aptitudes or skills (Goffard, 1960). Goffard said

> *For some men code skill seems to be impossible to learn, while for others it presents no problem. Although methods of selecting men with a high aptitude for learning IMC have been the object of research for a number of years, they are still only moderately satisfactory. With the least apt men eliminated by the Army Radio Code Test of the Army Classification Battery, the range of aptitude for code among men in courses which include IMC is still wide. p. 3.*

In one of the first studies of its kind, Thurstone (1919) used "mental tests" in an attempt to predict ability in telegraphy. He concluded that

> *The general intelligence tests are not as valuable for diagnosing ability to learn telegraphy as for measuring general intelligence. Ability in telegraphy is probably a special ability.... The fact that years of schooling does not agree with ability to learn telegraphy indicates that this is a special ability. College graduates usually do better on general intelligence tests than those who have only finished grammar school. But college graduates do not necessarily excel in learning telegraphy. p. 117.*

Low correlations have repeatedly been found between Morse Code achievement and intelligence, educational level, mechanical ability, and knowledge of subject matter. Woehlke (1956) summarized the research on attempts to identify selection devices for Morse trainees. He reviewed special tests of Morse aptitude including the "code learning" types of tests such as tests of code discrimination, learning, and speed of response. He found that attempts to link general ability, achievement, aptitude tests as well as non-test factors such as age and sex to code ability were unfruitful. Specific aptitude tests included auditory factor tests and clerical, musical and mechanical aptitude tests as well as many others. When compared to other areas of testing, Woehlke concluded that Morse code aptitude validities are inadequate.

From a dissertation to be submitted to the graduate school, University of Maryland, by the Author in partial fulfillment of the requirements for the Ph.D. degree.

The present attrition rate among Morse trainees is high. Attrition ranges from 26 to 42% of which "most" (i.e., 20-30%) is due to academic failure. Of the academic failures, nearly 100% are because of the inability of the students to copy code sent at a speed of 20 GPM (CODE3) -- the training standard. Historically, attrition has been reported to range from 18 to 60% in both military code training and in the "early" days of training railroad telegraphers (Woehlke, 1956). The Federal Communications Commission indicated that during the last 6 months of 1956, approximately 38% of the applicants for the General Class amateur radio operators license were rejected for failure to pass the code receiving test (Porter, 1957). The latter group would have possessed the motivation sometimes felt lacking in the military trainee.

Attempts to adjust the training or teaching methods have had only limited success in reducing the total number of hours required to train a Morse code operator and have not impacted significantly on reducing attrition. "Lengthening" the training period to reduce the attrition rate enabled more students to complete the school, but often these same students were rated unsatisfactorily by their supervisors on the job.

Compressed Speech

Time-compressed speech is defined (Foulke and Sticht, 1969) as speech which has been reproduced in less than the original production time. Compression is most frequently accomplished by electromechanically abutting periodic samples of the original recording. The end product is a tape with an accelerated word rate and a minimum of the frequency distortion associated with simple temporal alteration (e.g., playing a 33 1/3 rpm record at 78 rpm's).

Large individual differences with regard to the ability to comprehend compressed speech have been observed in the literature but have virtually been ignored in previous research.[1] This is understandable since most of the previous research has emphasized the use of compressed speech as a communications or educational medium in which individual differences were interpreted as error variance. The ability to comprehend compressed speech does not seem to improve significantly with listening exposure to compressed voice tapes. Training designed to improve comprehension of compressed speech has frequently been found to show little or no significant differences over neophite listeners (e.g., Foulke and Sticht, 1969).

Morse code can be thought of as the first language to undergo rate compression. It is unique when compared to compressed speech in that up to a rate of 20 GPM (CODE3) the integrity of the Morse code characters themselves does not change. That is, in compressed speech, small bits of words are usually randomly discarded in order to increase the word rate, thereby decreasing somewhat the intelligibility of the words. In speeding Morse code rate up to 20 GPM, the character sound is not changed, only the spacing between characters and groups is decreased to increase the rate. Therefore, the ability to comprehend time-compressed speech can be interpreted as "aptitude" related to the speed with which one can accurately process auditory stimuli. If so, it may provide an efficient technique to identify students for Morse code training.

[1] For a thorough review of the compressed speech literature, see Duker (1974).

METHOD

Subjects

The experiment was planned to include 120 service members enrolled in Morse code training at three Service schools. At the time of this analysis, data for 92 students was available. The students ranged in age from 17 to 31 (average = 20.5). Service grade ranged from E1 to E4 (average rank = E2). All subjects had normal hearing bilaterally as determined by a pure-tone audiometric screening test administered as part of the service selection battery for Morse operators. Each student had achieved an acceptable score on the Radio Code subtest of the Armed Services Vocational Aptitude Battery (ASVAB).

Test Instruments

Portions of the STEP (Sequential Tests of Educational Progress) Listening test were recorded and compressed[2,3] to create four listening comprehension test audio tapes (A, B, C and D) of equal length and difficulty. Each test contained three selections whose content ranged from junior high through high school level in difficulty. The recording time for each selection ranged from 62 to 91 seconds at an average rate of 187 words per minute (WPM). Each selection was followed by five multiple choice items. All items and the four response choices were read on tape. The response choices were also printed on the student answer sheets. Therefore, tests A, B, C and D each contained 15 items.

The selections in tapes B, C and D were time compressed at 1.5, 2 and 2.5 times normal, respectively. The test questions and distractors were not compressed. Thus, four levels of compressed speech (normal, 1.5, 2.0 and 2.5 times normal) were available as the repeated measures dimension in the experiment. The compressed tapes are referred to as BX, CX and DX and the WPM rates for each is 283, 340 and 434, respectively. Embedded in these tapes are tests B, C and D which are not compressed.

A "questions only" tape was prepared to assess the prose dependency of the test items. All tapes were presented using a Califone Model 3530 cassette recorder and MPC Model MX-200 headset.

Procedures

The experimental design incorporated two control and two experimental groups. For further control, all subjects were tested individually by a test proctor. An introductory tape was used to present test instructions. Further instructions were read to the S_s by the proctors. The presentation order of the tapes changed for each student to control for possible practice

[2] Reproduced by permission from Educational Testing Service, Princeton, N.J.

[3] Recordings were made at the Center for Rate-Controlled Recordings, University of Louisville.

effects. Control groups 1 and 2 (C_1 and C_2) were made up of twenty students each, who had just entered code school. Assignment was random. C_1 was administered test tapes A, B, C and D to assess the extent to which the four tests were equal in difficulty. C_2 was the prose-dependency control group. These students answered the same questions as the other groups in the experiment, but without benefit of listening to stories on which the items were based.

The experimental groups were administered tapes A, BX, CX and DX. Experimental group 1 (E_1) was made up of 30 randomly selected trainees who had met the 20 GPM (CODE8) code copying criterion. E_2 was comprised of 22 students who were being dropped from training because they were not making adequate progress toward achieving the criterion of 20 GPM.

After each experimental group student listened to the four tapes and answered test items, they listened to selections from tests BX, CX and DX again in order to assess speech intelligibility at the three compression levels. The students made a judgment on a scale of 0-100% to estimate what percentage of words from the selection that they thought they heard (independent of how they thought they performed on the earlier tests).

RESULTS

A univariate t-test compared the experimental groups on the sum of the four tests to see if there was an overall effect on the tests without regard to repeated measures. The difference was not significant ($t = 1.22$; $df = 1,50$; $p = .23$). A second test of interest was a one sample multivariate Hotelling T^2 to answer the question of whether there is a trend over the repeated measures dimension (levels A, BX, CX and DX). That is, do the scores drop off as speed increases? To do this, linear, quadratic and cubic contrast based scores were computed. The scores were transformed scores for each individual. The result of the T^2 test comparing A, BX, CX, and DX for E_1 and E_2 (combined) was significant ($T^2 = 132.73$; $df = 3,49$; $p < .001$). The relationship of the S_s test performance (comprehension) to rate of compression is best expressed as a straight line or linear trend ($t = -11.15$; $df = 1,51$; $p < .01$). The quadratic and cubic trends were not signficiantly better over and above linearity.

Figure 1 shows the means of the two experimental groups across levels of compression on each test. A two sample Hotelling T^2 was computed to test the interaction between the experimental groups across the repeated measures dimension. Once again transformed scores were used and contrast scores for the four treatment levels were computed to represent linear, quadratic and cubic trends. The results of the T^2 approached significance ($T^2 = 7.35$; $df = 3,48$; $p = .08$) but the null hypothesis that successful Morse trainess (E_1) and failures (E_2) do not differ, on the average, in their centroids on the comprehension tests compressed at three levels could not be rejected. Using the multivariate T^2 as an omnibus test to control for family-wise testing, similar to the Fisher lsd approach, would dictate that the analysis stop here. The decision would be that there is no significant interaction between E_1 and E_2 across compression levels. A less conservative approach, however, is to interpret the univariate t-tests based on the trend contrasts

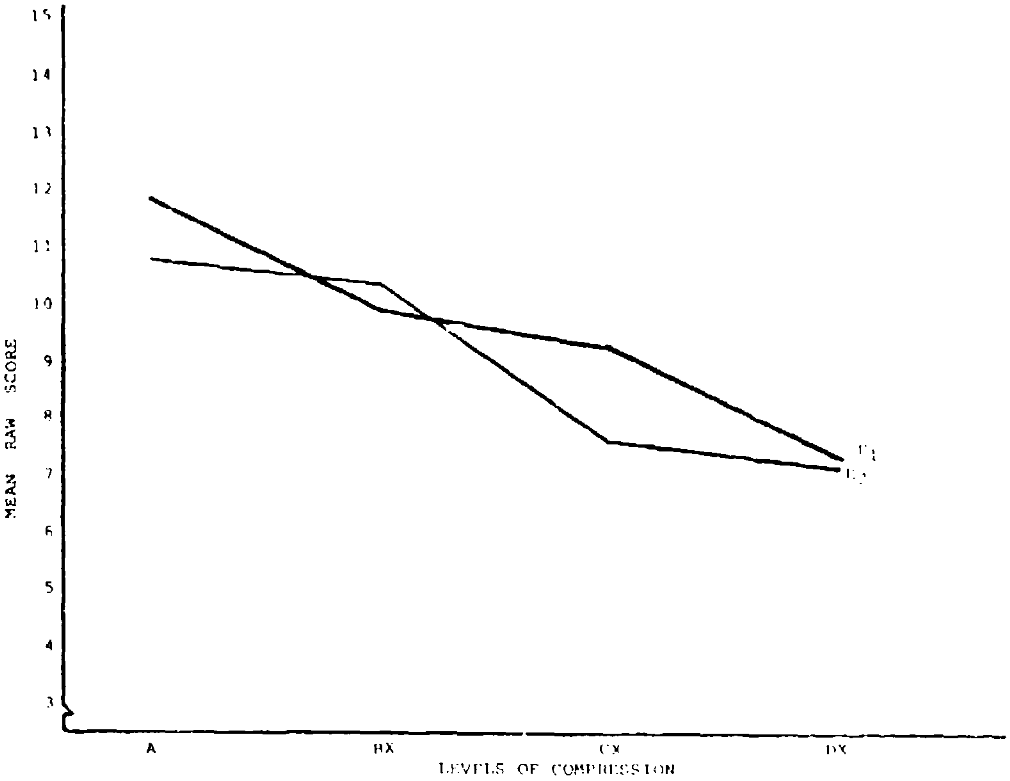


FIGURE 1. Average Test Scores Across Levels of Compression for the Experimental Groups

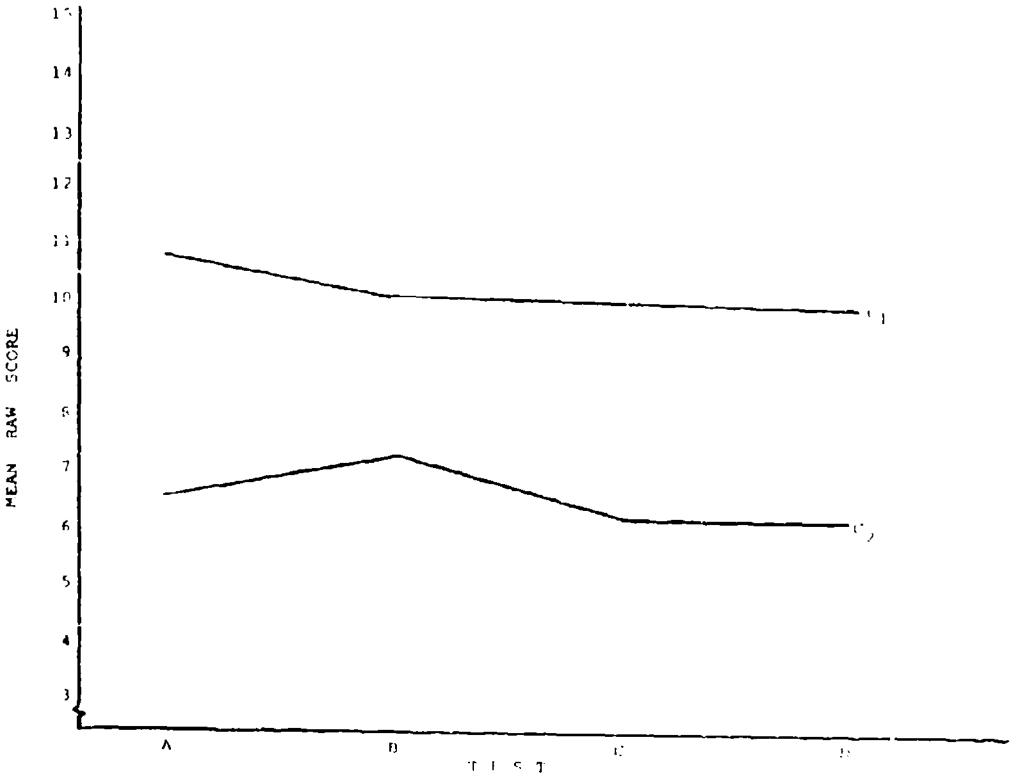


FIGURE 2. Average Test Scores for the Control Groups.

which are independent of the multivarate T^2. This analysis reveals a significant cubic trend ($t = -2.65$; $df = 1,50$; $p = .01$) and means that two significantly different curved (cubic) lines express an interaction across levels by group (see Figure 1). For reasons to be presented in the discussion, the latter analysis is preferred.

A univariate t-test showed that operators who achieve 20 GPM score significantly higher on level CX than did the failures ($t = 2.06$; $df = 1,50$; $p < .05$). No differences were found between the experimental groups on the other levels. One-way ANOVA's were computed to measure whether significant mean differences exist between groups on each test and on the sum of all tests. Significant F's, each with $p < .001$ and $df = 3,88$, were obtained for all tests. The Newman-Keuls procedure was run to determine where the differences occurred. On tests A and B group C_2 was significantly lower than E_1, E_2 and C_1, which were not different from one another. C_1 was significantly higher than C_2, E_1 and E_2 on test D. Again the latter groups were not significantly different from each other. Of most interest is the analysis of test C where E_1 was not significantly below C_1 and E_2 did not differ from C_2. However, E_1 and C_1 were significantly above E_2 and C_2.

An analysis of the sum of the test scores revealed that C_2 differs from all the other groups and leads to the intransitive decision that E_1 and E_2 do not differ, nor do E_1 and C_1, but that E_2 is significantly lower than C_1.

A one sample T^2 was used to test the hypothesis that performance on tests A, B, C and D were equal in the no-prose control group (C_2). Another test checked the equivalence of the difficulty level for the four tests for C_1 where only the normal level was presented. The null hypothesis was not rejected (as predicted) in either analysis indicating that the deviations in mean values of each test were not significantly different. In other words, the lines in Figure 2 representing C_1 and C_2 are not significantly different form horizontal lines (see Figure 2). The T^2 values were 1.88 ($df = 3,17$; $p = .65$) and 4.64 ($df = 3,17$; $p = .28$) for C_1 and C_2, respectively.

As shown in Figure 3, student perceived intelligibility of the prose also decreased significantly as a function of speed (see Figure 3). The one sample T^2 was equal to 359.78 ($df = 2,50$; $p < .001$). The univariate t for linear trend was significant ($p < .01$); however, the quadratic trend provided a significant improvement over and above linearity ($p < .001$). Therefore, a curved (quadratic) line best represents the relationship between perceived intelligibility and speed. The two sample mulivariate test for interaction between E_1 and E_2 across levels was not significant ($T^2 = 1.64$; $df = 2,49$; $p = .45$). The independent trend tests were also non-significant.

The correlation for the experimental groups between final code speed achieved in the course and the Radio Code (RC) subtest of ASVAB was .35 ($df = 1,45$; $p < .05$). Test C for the experimental groups did not correlate significantly with code speed ($r = .24$; $df = 1,50$; $p > .05$). Additionally, RC and C did not correlate with one another ($r = -.01$; $df = 1,46$; $p > .05$).

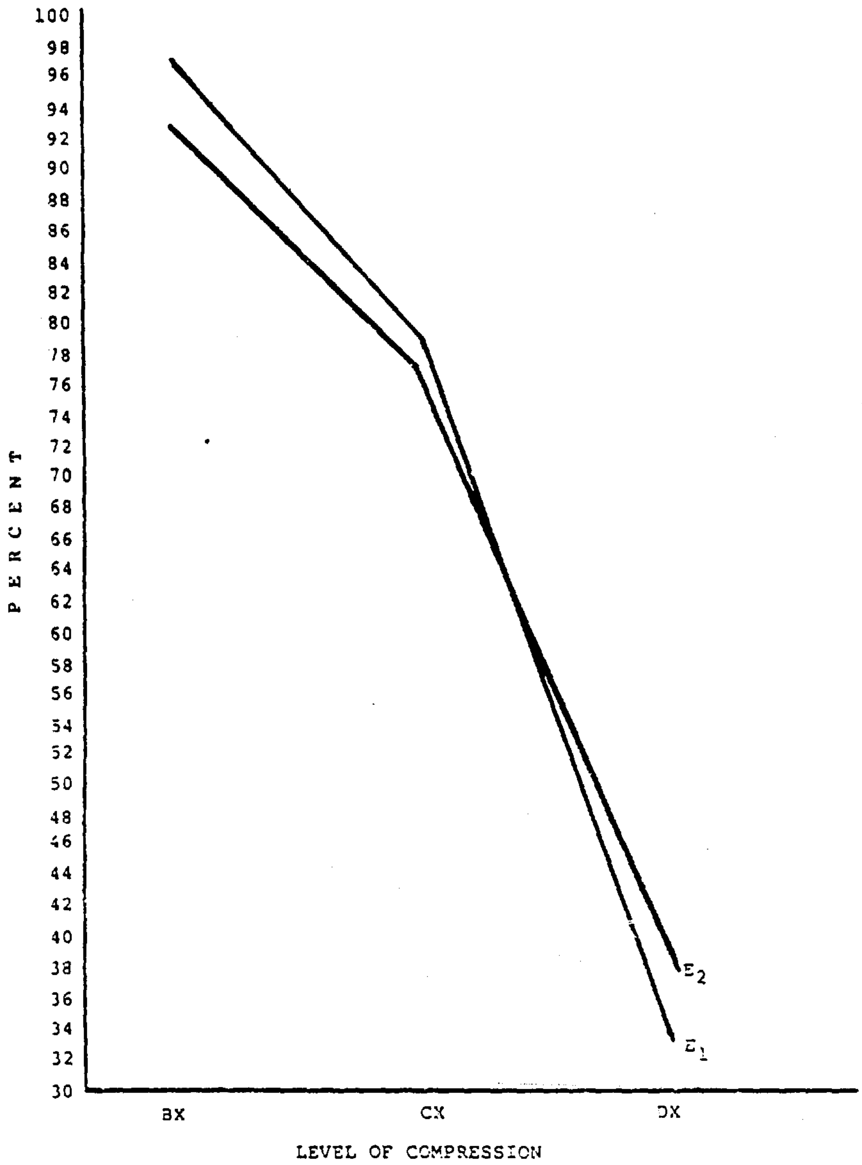


FIGURE 3. Group Mean Ratings of Intelligibility Across Levels of Compression

DISCUSSION AND CONCLUSIONS

There are several practical as well as theoretical inferences which can be based on the results of this study. Listening comprehension scores at normal speed did not discriminate successful code students from those who failed, supporting earlier findings as reviewed by Woehlke (1956). However, at word rates of 345 WPM (twice normal), successful code trainees were less seriously hampered in understanding the context of the prose selections than were the group of students who failed to meet the code speed requirements. In fact, the successful trainees' performance was not significantly different from the control group that listened to the uncompressed tapes. Furthermore, the unsuccessful students' test scores were not significantly different from the control group that did not hear the prose selections. The study has shown that the use of compressed speech has potential as a screening technique for selecting Morse Code trainees. It would certainly seem to warrant the expense of further research to develop a specialized test of compressed speech for this purpose.

The four tests used in the study can be regarded as equivalent based on the performance of the control groups. However, the group mean performance scores for C_2 were well above chance scores, indicating high information load (general knowledge) within the test items. In effect, the extent to which the tests were not prose dependent reduces their usefulness in studying the relationship of comprehension to compression. Because of the high information load of the test questions, rather than having 4 tests each with 15 items, the tests in effect have 8 or 9 items. To discriminate between the experimental groups using tests this short is difficult. Therefore, the less conservative statistical approach in interpreting the group differences seems warranted.

In order to make these findings more conclusive and to have a test of practical value in selecting recruits for code school, a longer comprehension test using prose selections of speeds ranging from approximately 320 to 380 WPM should be developed. The longer test then should be put to empirical scrutiny to assess the utility of this method.

Student perceived intelligibility of compressed speech was a function of speed. However, this technique of measuring the effect of compression was not useful for discriminating between successful and unsuccessful trainees.

The current selection device (RC subtest of ASVAB) for Morse training is a significant predictor of the final code speed students achieved. The lack of correlation between RC and test scores at level CX raises the possibility that two tests might be used in combination (multiple correlation) to produce a better selection procedure than could be obtained using either test singularly.

Further research using the ability to comprehend compressed speech as an aptitude may help in selecting students for other jobs which require auditory information processing, i.e., foreign language training. Individuals who are capable of processing auditory stimuli more rapidly (or have larger auditory channel capacity) may be more apt in carrying out the cognitive functions of "translating" the second language back to their first while auditory input continues.

REFERENCES

Duker, Sam. Time-Compressed Speech: An Anthology and Bibliography in Three Volumes. New Jersey: The Scarecrow Press, Inc., Volumes I, II, and III, 1974.

Foulke, Emerson; and Sticht, Thomas G. "Review of Research on the Intelligibility and Comprehension of Accelerated Speech." Psychological Bulletin, 1969, Vol. 72(1), pp. 50-62.

Foulke, Emerson; and Sticht, Thomas G. "The Intelligibility and Comprehension of Time Compressed Speech." Proceedings of the Louisville Conference on Rate-Controlled Speech, distributed by American Foundation for the Blind, 1975, pp. 21-28.

Goffard, S.J. Experimental Studies of Skill in Copying International Morse Code." HUMRRO Technical Report 68, December, 1960.

Porter, Charles Baddeley. An Experimental Investigation of Selected Variables Related to Morse Code Learning. Unpublished doctoral dissertation, University of Illinois, 1957.

Thurstone, L.L. "Mental Tests for Perspective Telegraphers, A Study of the Diagnostic Value of Mental Tests for Predicting Ability to Learn Telegraphy." Journal of Applied Psychology, 1919, Vol. 3, pp. 110-117.

Thurstone, L.L. "The Selection and Training of Telegraphers." Psychological Bulletin, Feb, 1919, Vol. 16, pp. 58-59.

Woehlke, Arnold Benjamin. "The Construction and Evaluation of the International Morse Code Selection Test." Unpublished doctoral dissertation, Boston University, School of Education, 1956.

THE EFFECTS OF A CORRECTION TO THE OWEN ALGORITHM

Steven Gorman

Navy Personnel Research and Development Center
Washington Liaison Office

INTRODUCTION

The vast changes in computer technology have made a strong impact upon the field of ability measurement. The increased capabilities and decreased costs of computer use have opened the door to application of latent trait (also called item characteristic curve) theory. As a result, this has made possible the psychometrician's dream of individualizing tests for each examinee.

Two Bayesian procedures for ability estimation have become popular, the Owen (1975) Algorithm and the Bayes modal procedure (Samejima, 1969). The Owen Algorithm has been researched only in the adaptive mode, with the exception of one study by this author (Gorman, 1979). The Bayes modal procedure has only been used in the static mode. This paper will investigate the use of this procedure in an adaptive mode.

McBride & Weiss (1976) studied the Owen Bayesian adaptive procedure and determined that with this procedure, test scores regress toward the mean. That is, high ability examinees tend to have lower ability estimates. Urry (1977) has reviewed this study, and has suggested a correction, namely dividing the Bayesian regressed ability estimate by the test reliability. A second, potentially more serious problem is the reliance upon accurate three parameter logistic item parameters. Urry (1976) developed OGIVIA3, a computer program to estimate these required values. The effectiveness of this estimation procedure for use in the Owen Algorithm was reviewed by Gugel et. al. (1976). OGIVIA3 has been revised (Croll & Urry, in preparation) and is named ANCILLES.

The purpose of the present paper is to evaluate the effectiveness of two Bayesian adaptive ability estimation procedures with a correction for regression using known item parameters and those estimated by ANCILLES. The results will be compared to an "ideal" test consisting of test items of the same quality, and scored both with the conventional raw score to Z transformation, and with Bayes modal scoring. The issues investigated are a) correlation of estimated ability with true ability, b) computer processing time, c) conditional bias, d) conditional accuracy and e) test score precision.

The Owen Algorithm

The Owen Bayesian adaptive ability estimation procedure has been well documented elsewhere (Owen, 1975; McBride and Weiss, 1976) and will not be reported here. However, to comprehend the correction, a brief conceptual description is in order. The procedure assumes a normal distribution with mean 0 and variance 1. The item bank is then scanned to determine which item will minimize the expectation of the posterior variance of the distribution if administered. The item is then administered and a new ability estimate (mean of posterior distribution) and variance about that estimate are computed. The item characteristic (also called item response) function is employed as the likelihood function. The product of the prior distribution and the likelihood function is the posterior variance of ability. The ability estimate is then used as the prior mean, and an item is again selected to minimize the expected value of the posterior distribution variance. This procedure is repeated iteratively until a specified value of posterior variance or number of items is attained. A correction for regression is applied to the final ability estimate. The correction consists of dividing the final ability estimate by what Urry (1977) refers to as the test reliability. This reliability is the value unity minus the Bayesian posterior variance, and this value obviously will differ for each individualized test. Urry believes that more accurate measurement is attained by terminating adaptive tests based on a fixed posterior variance, rather than a fixed number of items. However, he concedes that this correction should be effective for both fixed and variable length tests. This study pursues the fixed length only.

The Bayes Modal Procedure

The Bayes modal adaptive ability estimation procedure developed for this study consists of two algorithms, one to estimate ability, and one to select appropriate items to be administered to each examinee. The ability estimation algorithm is based on the Bayesian scoring procedure developed by Samejima (1969), based on the item response function and an assumption of a normal distribution of ability. Urry (1976) incorporates this procedure into the second iterative stage of his item parameter estimation procedure. The item selection procedure chooses that item which provides the most item information for the current ability estimate. The item response function for all administered items is computed. The product of all administered items is computed. The product of all item response functions and the assumed normal density function is the posterior distribution, the mode of which is the ability level estimate. This mode is selected through the use of the iterative Newton-Raphson Algorithm. This final ability estimate is then unregressed by dividing by the validity, the square of the reliability, for that set of administered items. This larger correction is required because this adaptive procedure yields ability estimates which are more regressed toward the mean than the Owen ability estimates. The Bayes modal scoring of static tests consists of the scoring algorithm (described above) employed on the entire test and corrected for regression by dividing by the test reliability (described in the Owen Algorithm paragraph).

Method

To investigate the criteria mentioned above, artificial data were generated according to the three-parameter logistic model:

$$P_i(\theta) = c_i + (1-c_i)[1 + \exp(-1.7a_i(\theta-b_i))]^{-1} \qquad (I)$$

where:

$P_i(\theta)$ = Probability of a correct response
a_i = item discriminatory power
b_i = item difficulty
c_i = item coefficient of guessing

using the LVGEN program developed by Urry (1971). This program provided vectors of responses, correct (1) or incorrect (0) for the simulated examinees (sims).

An "ideal" bank was created consisting of 101 items at equal increments of b_i = .05 over the range -2.5 to +2.5. This bank used items whose item discriminatory power was set at a_i=1.6, and item coefficient of guessing c_i=.15. These values represent attainable values for high quality multiple choice items. A thirty item rectangular test was developed from this high quality level of items, with item difficulty (b_i) values spread evenly over the interval -2.5 to +2.5.

These test items were included in a bank of items from another study (Gorman, in preparation) making a total of 873 items with a_i values at three levels. The item parameters were estimated by the ANCILLES program based on the responses of 2000 simulated examinees (sims) on tests of 51 item length.

Two populations of sims were used in this study. A group of 500 sims representing a normal population were generated from the LLRANDOM program (Learmonth & Lewis, 1973). This population was used to determine the fidelity coefficient, the correlation of known with estimated ability, and computer processing unit (CPU) time. The second population consisted of 100 examinees at each of 11 equally spaced ability levels on the interval -2.5 to +2.5. The static and adaptive tests were simulated using the known and estimated item parameters. With the normally distributed population of 500 sims, the following criteria were evaluated.

a) Fidelity Coefficient

A Pearson product-moment correlation between known and estimated examinee ability was computed for all tests.

b) Computer Processing Time

The amount of computer processing unit (CPU) time for the two adaptive tests were recorded.

With the rectangular distribution of examinees, the following additional criteria were evaluated:

c) Conditional bias,

This statistic provided an indicator of the magnitude and direction of the error between true ability and ability estimated by each of the test procedures at various levels of the trait continuum

$$\text{bias} = b_e | \theta_e = \hat{\theta}_e - \theta_e \qquad (2)$$

where:

b_e = bias values at each ability level

$\bar{\hat{\theta}}_e$ = average ability estimates at each ability level

θ_e = true ability of examinees at each ability level

d) Conditional Accuracy,

The accuracy of the test scores will be provided by the root mean square error computed at the 11 ability levels by

$$e_i|\theta = [\Sigma(\hat{\theta} - \theta)^2 \quad n^{-2}] \tag{3}$$

where:

$e_i|\theta$ = root mean square error conditional upon ability level

n = 100

θ = known ability level

$\hat{\theta}$ = ability estimate

e) Test Score Precision

Test score precision will be provided by the test score information function:

$$I_x(\theta) = \left[\frac{\delta/\delta\theta(E(x|\theta)}{\sigma_{x|\theta}}\right]^2 \tag{4}$$

where:

x is a set of test scores

$E(x|\theta)$ is the expected value of the test scores at θ

$\sigma_{x|\theta}$ is the variance of test scores at θ

For all static and adaptive tests, the test score information function was approximated by:

$$I(_{\hat{\theta}}(\theta') = \left[\frac{\delta/\delta\theta \;\; E(\hat{\theta}|\theta')}{\sigma_{\hat{\theta}|\theta'}}\right]^2$$

where:

$I_{\hat{\theta}}(\theta')$ is the test score information at 9 trait values

$E(\hat{\theta}|\theta')$ is the slope of a curve fit to three consecutive test score means $(\theta'-.5, \theta', \theta'+.5)$

$\sigma_{\hat{\theta}|\theta'}$ is the variance of test scores at θ'

Results

Fidelity Coefficient

The Pearson product-moment correlations between known and estimated ability for the static and adaptive tests are in Table 1.

TABLE 1

Fidelity Coefficients Using Known and Estimated Item Parameters

	STATIC TEST		ADAPTIVE TESTS	
SCORING	ITEM PARAMETERS		BAYES MODAL	OWEN
Bayesian	Known	.911	.943	.945
Bayesian	Estimated	.907	.940	.943
Z-Score		.894		

The differences between Bayesian and conventional scoring of the rectangular test, and between either scoring of the static test and the adaptive tests are both statistically and practically significant.

Computer Processing Time

Computer processing unit (CPU) time required to administer 30 items to 500 sims on the IBM 360/165 computer was 157 seconds for the Bayes modal strategy and 233 seconds for the Owen procedure.

Estimation bias

Table 2 lists the estimation bias as a function of ability for the static and adaptive tests.

TABLE 2

Bias as a Function of Ability and Known or Estimated Item Parameters

	STATIC TEST			ADAPTIVE TESTS			
		BAYES MODAL		BAYES MODAL		OWEN	
ABILITY	Z	KNOWN	EST	KNOWN	EST	KNOWN	EST
-2.5	.351	.136	.271	-.054	.103	-.017	.081
-2.0	.171	.068	.159	-.010	.070	-.076	.008
-1.5	.119	.078	.133	.081	.139	-.013	.043
-1.0	.196	.108	.162	.029	.071	-.058	.000
-0.5	.096	.031	.080	-.008	.003	-.055	-.033
0.0	.065	.027	.062	.019	.031	.013	.030
0.5	.035	-.008	.013	.033	.058	.066	.121
1.0	-.021	-.053	-.027	-.028	-.002	.040	.059
1.5	-.012	-.032	-.011	.012	-.003	.125	.075
2.0	-.099	-.071	-.173	-.001	-.184	.147	.047
2.5	-.320	-.153	-.366	.014	-.215	.162	.042

The table shows that for all three conditions of the static test, and for the Bayes modal adaptive test with estimated item parameters, curvilinear bias is evident. It also shows the adaptive tests providing less biased ability estimates than the static test.

Estimation accuracy

Table 3 lists the estimation accuracy (root mean square error) as a function of ability for the static and adaptive tests.

TABLE 3
Root Mean Square Error* as a Function of Ability
and Known or Estimated Item Parameters

	STATIC TEST			ADAPTIVE TESTS			
		BAYES MODAL		BAYES MODAL		OWEN	
ABILITY	Z	KNOWN	EST	KNOWN	EST	KNOWN	EST
-2.5	521	374	416	308	250	230	257
-2.0	402	318	321	190	239	236	216
-1.5	430	311	336	208	230	204	237
-1.0	442	308	328	164	166	184	165
-0.5	375	323	324	202	191	228	210
0.0	349	289	278	182	190	202	208
0.5	355	294	289	189	193	220	221
1.0	284	278	298	181	204	200	204
1.5	295	312	293	183	158	243	191
2.0	249	316	286	198	252	251	214
2.5	365	339	473	224	359	273	240

* Decimal points omitted

Table 3 indicates that the adaptive tests provide more accurate ability estimation than the static test. The Bayesian scoring of the static test, even with fallible item parameter estimates, is generally more accurate than the conventional scoring. All tests were more accurate about the mean than at the extremes of ability. This trend is more evident on the static test and the Bayes modal adaptive test than the Owen adaptive test.

Test Score Precision

Test score information as a function of ability for the static and adaptive tests are listed in Table 4. The Bayesian scoring of the adaptive test provides its greatest benefit over conventional scoring at the low ability levels. The two adaptive tests provide three times the level of precision of the static test. Additionally, the adaptive tests yield roughly equal test score precision over the entire ability range investigated.

TABLE 4
Information as a Function of Ability and
Known or Estimated Item Parameter

	STATIC TEST			ADAPTIVE TESTS			
		BAYES MODAL		BAYES MODAL		OWEN	
ABILITY	Z	KNOWN	EST	KNOWN	EST	KNOWN	EST
-2.0	4.38	9.09	9.39	33.97	19.85	19.37	19.54
-1.5	5.71	11.76	10.35	30.04	30.55	25.57	18.50
-1.0	6.31	11.26	11.23	31.93	33.88	30.03	31.67
-0.5	4.91	7.72	7.12	23.10	23.56	22.82	23.33
0.0	7.49	11.16	11.87	32.90	31.79	30.90	31.58
0.5	7.68	10.38	11.37	27.18	29.44	24.65	32.81
0.0	11.69	13.12	11.06	31.07	22.08	30.48	25.31
1.5	9.57	9.86	8.19	31.11	24.99	27.66	31.59
2.0	8.99	8.10	8.07	25.34	22.02	25.91	21.60

Discussion

The Owen adaptive test performs slightly better than the Bayes modal adaptive test on the bias criteria, and equal on the accuracy precision, and fidelity coefficient criteria. The one criterion which separates the two tests is the amount of computer processing time required. In this case, the Bayes modal adaptive test is clearly superior, since the Owen procedure requires 50% more CPU time. This may be due to the subroutine which selects the most appropriate item for the Owen. It selects that item which minimizes the expected value of the posterior variance, and involves computing a quadratic loss function for all items not yet administered to the examinee in the item bank. The Bayes modal adaptive procedure uses a simpler mathematical formula to determine the most appropriate item after each step. Both adaptive procedures could be made for time-efficient by using a lock-up table procedure, which could be be created prior to test administration. This could possibly result in a slightly less appropriate item selection, but this needs to be researched. One impressive point exhibited by these data is how robust the two Bayesian adaptive and static tests are to the use of fallible item parameters. The Bayes modal adaptive test, even with the correction for regression, gives regressed ability estimates. One simple solution might be to score the final test results with maximum likelihood scoring.

The fallible item parameters generated by ANCILLES are used in this study to show the decrement in measurement properties that we would expect to find with live testing. The results seem to be within limits that we can live with, especially when compared to an ideal static test. The inverse of the square of the test score information can be thought of in terms of the local standard error of estimate. The adaptive tests reduce the standard error of estimate to what we would expect from a static test three times as long (90 items). The Bayesian scoring of the static test is also impressive when compared to the conventional scoring, especially at the low ends of ability, where guessing is more likely to be effective.

Summary

The two Bayesian adaptive tests investigated provide relatively equal measurement of ability. The Bayes modal procedure was much more efficient in terms of computer processing time. Both adaptive tests gave as much precision as a static test three times the length. Also, Bayesian scoring of the static test gave better measurement properties than conventional scoring.

References

Croll, P. and Urry, V.W. *ANCILLES: A program for estimation of the item parameters of normal ogive and logistic mental test models.* U.S. Civil Service Commission, Washington, D.C., in preparation.

Gorman, S. *A comparison of Bayesian adaptive and static tests using a correction for regression.* Paper presented at the Third Biennial Conference on Computerized Adaptive Testing, Wayzata, Minnesota, June, 1979.

Gorman, S. *A comparative evaluation of two Bayesian adaptive ability estimation procedures with a conventional test strategy.* Doctoral dissertation, Catholic University, in preparation.

Learmonth, G. E. and Lewis, P. A. W. *Naval Postgraduate School Random Number Generator Package: LLRANDOM.* Research Report NPS55LW73061A, Naval Postgraduate School, Monterey, CA., 1973

McBride, J. R. and Weiss, D. J. *Some properties of a Bayesian adaptive ability testing strategy.* Research Report 76-1. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 1975, *70* (350), pp. 351-356.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph* No. 17, 1969.

Urry, V. W. A Monte-Carlo investigation of logistic model test models. Doctoral dissertation, Purdue University, 1970, *Dissertation Abstracts Interational*, 1971, *31*, 6319B. (University Microfilms No. 71-9475).

Urry, V. W. Ancillary Estimators for the Item Parameters of Mental Test Models. In *Computers and Testing: Steps Toward the Inevitable Conquest*, Professional Series 76-1, Washington, D. C.: Personnel Research and Development Center, U. S. Civil Service Commission, 1976.

Urry, V. W. *Tailored testing: A spectacular success for latent trait theory.* (TS-77-2), Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, 1977.

AN EVALUATION OF COMPUTERIZED ADAPTIVE TESTING

James R. McBride

Navy Personnel Research and Development Center

INTRODUCTION

Problem

Since January, 1976, all military services have used a common battery of mental tests for enlisted personnel selectiona and classification: the Armed Services Vocational Aptitude Battery (ASVAB). The battery includes twelve subtests of cognitive aptitudes. These subtests are necessarily short; they are usually scored by hand; raw scores are manually converted into service-specific scaled scores using conversion tables; scale scores are manually recorded, and manually transcribed into permanent individual personnel records.

The U.S. Marine Corps has perceived some difficulties with the ASVAB testing program. Now that ASVAB has supplanted service-specific classification test batteries, a single test battery must serve all the special testing needs of the four services. In many cases, ASVAB subtests are excessively difficult for Marine Corps selection and classification purposes; this can result in inefficient and inaccurate classification. There has been some compromise of ASVAB test security--test booklets and answer keys have been stolen; this problem , if uncontrolled, could seriously degrade the validity of the tests for classification purposes. The manual nature of the test scoring, score conversion, and score recording procedures provides opportunity for clerical error, and it is believed that such errors may have resulted in numerous accession errors.

The Marine Corps formulated an operational requirement to lessen or eliminate the impact of the problems discussed above. Computer-administered adaptive testing (CAT) was identified as one potential solution to all of these problems. In an adaptive test, test difficulty is tailored dynamically to the ability level of the individual examinee; in principle, then, CAT eliminates the problem of excessive test difficulty, and should yield scores which promote accurate selection and classification decisions. CAT addresses the test security problem by eliminating printed booklets and scoring keys, and by administering an individually tailred set of test items to each examinee. Additionally, since CAT automates test administration, test scoring and recording are automated as well, thereby eliminating human clerical error from the testing system.

Recognizing the potential of CAT for selection and classification testing, the Marine Corps tasked NPRDC with investigating the feasibility of CAT, as part of a program of phased research and development related to military personnel accessioning.

Purpose

The research reported here was intended to assess the feasibility of using computerized adaptive testing (CAT) in a Marine Corps recruit/ applicant population, and at the same time to verify the claimed merits of CAT as a psychological measurement technique. These two research issues could only be addressed by administering adaptive tests to appropriate examinee samples. The capability to to this had to be developed: equipment identified, software written, and large banks of test items assembled and calibrated using item characteristic curve models. After this development was completed, a pilot study involving verbal ability tests was conducted. This report describes the pilot study of the feasibility and psychometric merits of an adaptive procedure for measuring verbal ability.

Background

Group-administered paper-and-pencil "objective" ability tests date back to World War I, when the introduction of the Army Alpha test signalled an era of vast improvements in the administrative efficiency of psychological testing. The price paid for this efficiency was loss of flexibility, since all examinees must answer a common set of test questions. The psychometric effect of this was not too serious, provided that a test was designed to have a difficulty level appropriate to its intended application, or that a test was sufficiently long to overcome minor design deficiencies. For persons whose ability level was not near the target difficulty level of the test, however, the paper-and-pencil test was not a particularly accurate or precise measuring instrument.

The psychological tests used by the armed services for selection and classification are group-administered paper-and-pencil tests. Such tests, as just discussed, lack the flexibility to measure well over a wide range of ability. In order to achieve that flexibility, the difficulty level of the test would have to be chosen to fit individual ability levels. This is not practical, since individual ability levels are not known prior to to testing, but it can be accomplished using an adaptive test, in which test items are chosen sequentially on the basis of the examinee's performance. This sequential item choice can best be accomplished using automated test administration; for example, by having the test administered at an interactive computer terminal.

The historical development of computer-administered adaptive testing was reviewed by Weiss and Betz (1973) and by Wood (1973).

Weiss (1974) surveyed a variety of alternative adaptive testing methods; the same author (Weiss, 1975) summarized a number of potential advantages of CAT over conventional paper-and-pencil tests. Despite those advantages, most research into adaptive testing was done at the basic research level, until the U.S. Civil Service Commission began moving toward early 1980's implementation of computer-based adaptive adminstration of its PACE examination (Gorham, 1975).

The Civil Service Commission implementation plans were based on research conducted by Urry and his colleagues (e.g., Urry, 1977). Urry chose to adopt a Bayesian sequential adaptive testing procedure proposed by Owen (1969, 1975), and demonstrated that the procedure could achieve satisfactory levels of measurement reliability in substantially less than half the length required of a conventional test. In one instance, Urry estimated that an adaptive test was equivalent in reliability to a conventional test 5 times as long (Urry, 1977). It is this efficiency of measurement which has motivated most psychometric interest in adaptive testing, although test users have often been more attracted by its practical advantages, which were discussed above.

Marine Corps interest in CAT for personnel selection and classification testing resulted from dissatisfaction with certain aspects of the joint service paper-and-pencil testing battery. Subtests used for selection decisions were also used as a basis for personnel classification and assignment to specialized training; a test designed for one of these purposes would likely be inappropriate for the other, and this might result in disproportionate numbers of selection or assignment errors. Clerical errors in the manual scoring and score recording processes were felt to be another serious source of accessioning errors; and the effects of test compromise were inevitable with the use of the same test battery over a period of several years.

Recognizing that computerized test administration could eliminate scoring and clerical errors, and that adaptive testing could substantially reduce test compromise, Headquarters, Marine Corps, tasked NPRDC with evaluating the feasibility of CAT for testing Marine recruits. The purpose of this paper is to report the results of the first in a series of studies investigating both the feasibility and the utility of CAT in comparison with a conventional test design.

The study was designed in part to address these research questions: 1) Is computer-based testing of military recruits administratively feasible? 2) Is a computer-administered adaptive test more reliable than a conventional test, holding test length constant? 3) If so, what is an appropriate length criterion for an adaptive test?

These questions were motivated by the results of previous research done elsewhere. The first question, that of administrative feasibility, seems trivial, but is not. Interviews with military testing personnel indicated some misgivings about the ability of military recruits to use relatively sophisticated automated testing equipment, such as CRT computer terminals. This potential man-machine interface problem is the analogue of administrative difficulties encountered years earlier with paper-and-pencil tailored tests. For example, Seeley, Morton, and Anderson (1962) found that a substantial proportion of their military examinees did not successfully follow instructions on an experimental sequential item test; this experience may have caused a five-year lapse in military research on tailored or adaptive testing. Olivier (1974) had a similar experience using a paper-and-pencil flexilevel test in a sample of high school students.

The question of the advantages of adaptive tests over conventional ones in terms of reliability has a clear and positive theoretical answer: holding test length and all else constant, a good tailored test design is superior, provided that highly discriminating test items are available (Urry, 1970).

This theoretical advantage is not always corroborated in empirical investigations. For instance, Bryson (1971) questioned the advantage of tailored testing over certain methods of conventional test design; Olivier (1974) failed to find an advantage for the flexilevel tests he used; and the results reported by Weiss and his colleagues have been less than unanimous in favor of adaptive tests. All these results are in contrast with those of Urry (1977) who reported that for his sample of 57 Civil Service job applicants, an adaptive verbal ability test achieved an 80-percent reduction, compared to a conventional test, in the test length required to attain any of several specified levels of reliability. Urry's result was extraordinary. The only cloud over it is that it was based on indirect evidence: the conventional test reliabilities were based on Spearman-Brown equation adjustments to the reliability obtained in an independent sample; and the tailored test reliability was merely assumed, not rigorously verified .

Previous research into the reliability, validity and efficiency of adaptive tests has often been inconclusive because of design flaws or nuisance factors. The major problem has been the lack of suitable means for estimating the adaptive test's reliability without making dubious assumptions. Another problem has been the general failure to match adaptive and counterpart conventional tests in item quality, with an unfair advantage usually going to the adaptive test. The research reported here was designed intentionally to remove those two problems-- to provide credible indices of reliability which are appropriate for both test types, and to provide a fair comparison by matching item quality across the test types. With those two problem sources eliminated, there is hope for an unequivocal comparison between adaptive and conventional test designs.

APPROACH

The general method used was that of equivalent tests administered to independent examinee groups. One group took two equivalent computer-administered adadptive tests. The other group took two equivalent conventional tests, also administered by computer. In order to control for item quality, both types of test were made up of items from the same source--a common pool of 150 verbal ability items which had been previously calibrated, using item characteristic curve methods, in large samples of Marine recruits.

Research Design

Each examinee was randomly assigned to one of the two treatment groups: A or C . Group A took two 30-item adaptive verbal ability tests, followed by a 50-item criterion test of word knowledge. Group C took two 30-item conventional verbal ability tests, followed by the same criterion test. All tests were administered at a computer terminal. Figure 1 is a schematic representation of the research design.

Observations

For each examinee who completed the tests, the following data were observed and automatically recorded:

1) Elapsed time for the testing session;
2) Elapsed time to complete pre-test instructions;
3) The number of errors made during the instructions;
4) The number of times the proctor was called;
5) Raw item scores (correct/incorrect);
6) Cumulative raw score after each item;
7) Latent trait ability estimates (experimental tests only);
8) Bayes posterior variance of the ability estimate after each item;
9) Criterion test raw score.

The format for these observations is schematized in Figure 2.

Independent Variables

For the comparisons between the adaptive and conventional testing methods, there were two independent variables:

- Test type (adaptive vs. conventional);

- Test length (5, 10, 15, 20, 25, 30 items).

Within the adaptive testing method, the test ermination rule was treated as in independent variable for some analyses:

- Termination rule: Test length (5, 10, ... ,30 items)

  vs.

  Specified posterior variance (variable length).

The number-of-items termination rule results, of course, in a test of predetermined length; the posterior variance rule results in a variable length test, depending on the number of items required to attain specified levels of the Bayes posterior variance.

Dependent Variables

Measures of the dependent variables were formed from the individual observations. The dependent variables included:

-
- Testing time
- Instruction time
- Number of keyboard errors
- Number of proctor calls
- Alternate tests reliability after 5, 10, ..., 30 items
- Test-criterion correlation after 5, 10, ..., 30 items .

Items

The 150 items in the pool were calibrated using Urry's ancillary estimation method, and were selected according to the prescriptions given by Urry (1977): all item characteristic curve slope parameters exceeded .80 (the average value of this "a"-parameter was 1.24); item difficulty (location) parameters ranged from -2.0 to +2.0; no item had a pseudo-guessing ("c") parameter greater than .30.

Examinees

Male Marine recruits reporting for duty at the Marine Corps Recruit Depot, San Diego, were the examinees. They were tested one at a time at a Burroughs TD832 terminal controlled by a Burroughs B1717 time-sharing minicomputer system. Assignment to groups ("A" or "C") was randomized. 201 examinees completed the tests; 96 of these took the adaptive tests; 105 took conventional tests.

Tests

The conventional tests administered to Group C were "rectangular" tests spanning the difficulty range of the item pool. Their broad range of difficulty was chosen in order to simulate the psychometric design of the verbal tests used in the Armed Services Vocational Aptitude Battery. Two 30-item equivalent forms were constructed from the 150-item pool. Items were chosen to be as highly discriminating as possible, consistent with the broad difficulty range. The two forms were constructed to be "weakly parallel" (Samejima, 1977), i.e., to have approximately equal test information functions. Within each form, the 30 items were sorted into five difficulty levels, then arranged in descending order of discriminating power within each level. The first five items in each form were the most discriminating items at their respective difficulty levels; items 6 through 10 were the second most discriminating items at each level; and so on. This arrangement resulted intwo 30 item tests consisting of a sequence of six 5-item subsets each. This design was intended to permit meaningful analysis of the psychometric properties of rectangualr conventional tests of length 5, 10, 15, 20, 25, and 30 items. In order to equalize any effects due to test length, fatigue, or other extraneous factors, the two conventional tests were administered in counterbalanced item order; i.e., the two 30-item tests were administered as one 60-item test in the following order:

Item sequence:	1	2	3	4	5	6	7	8	...
Test Form:	1	2	2	1	2	1	1	2	...

The two 30-item adaptive tests were based on Owen's Bayesian sequential tailored testing procedure. For each examinee, and each test form, an initial normal prior distribution of ability was assumed, with mean 0 and variance 1.0 . Test order (1 or 2) was counterbalanced for each examinee in a manner identical to that of the conventional tests: 12212112... Both "forms" of the Bayesian tests drew items from the same 150-item pool; counterbalancing the order of administration here served the added purpose of equalizing item quality across the two forms. The two adaptive tests were independent of each other, except for their use of a common item pool.

The criterion "test" was formed by concatenating two obsolete operational test forms measuring word knowledge. This resulted in a 50-item test expected to be a highly reliable and fairly broad-range test of an important facet of verbal ability.

RESULTS AND DISCUSSION

Feasibility

Data pertaining to the feasibility of using computer terminals to administer tests to military recruits are summarized in Table 1.

Mean testing time was 61.0 minutes for the adaptive test group, versus 50.4 minutes for the conventional test group. These are the mean times to answer 110 items-- 60 items from either the adaptive or the conventional alternate forms, followed by 50 criterion test items common to both groups. After making a few assumptions, a little algebraic manipulation of these testing times yields the result that the average adaptive test took about 11 minutes longer than the average conventional test. The adaptive tests, then, required about 11 ,ore seconds per item, or as much as 39 .per cent longer to answer than the conventional tests. Some or all of this difference may have been due to computations required for adaptive item selection, but this result does agree generally with Waters' (1977) finding that an adaptive test required significantly longer examinee processing per item than a similarly administered conventional test. In the present study, however, the observed time difference may be due in a large part to idiosyncrasies of the computer system; if so, differences of the size reported here would not be expected if a faster scientific computer were used to control and administer the adaptive tests.

Instruction time averaged 9.5 minutes for the adaptive test group, and 10.3 minutes for the conventional group; overall, the instructions required an average of 9.9 minutes. During this time, the examinees were familiarized with the CRT and keyboard by means of a programmed instructional sequence with special branching following procedural errors, and with an audible call to the proctor if the examinee had difficulty correcting an error. Errors and proctor calls were counted. As the table indicates, there were 55 errors in all in 201 test sessions; in only 17 cases was the proctor called. This amounts to about one procedural error per four test sessions, and to a requirement for proctor intervention about one time per twelve test sessions.

Psychometric Characteristics

Table 2 summarizes reliability and criterion validity data for both the adaptive and conventional alternate forms tests, at lengths of 5, 10, 15, 20, 25, and 30 items.

Reliabikity is operationalized here as the correlation between scores on alternate forms at a given test length. The scoring procedure used was the same for both test types -- latent ability estimation using the sequential estimation formulae developed by Owen (1969). From the table, it is clear that the adaptive tests had substantially higher reliability coefficients than the conventional tests, for any given test length. Looking at these data another way, we see that the adaptest reliability at a 5-item test length was practically equivalent to the conventional tests' reliability at 15 items; similarly, the adpative tests' reliability at length 10 was superior to that of the conventional test at length 25.

Figure 3 contains a graphic comparison of the adaptive and conventional tests in terms of alternate test reliability as a function of test length. Analysis of Table 2 and Figure 3 indicates that in terms of test length required to attain a given level of reliability, the adaptive tests had a substantial advantage over the conventional tests. This advantage was essentially the same for both fixed length and variable length stopping rules; there was no apparent advantage to variable length, as opposed to fixed length, within the adaptive testing methdd.

Thus, the adaptive tests achieved specific levels of reliability more efficiently than the conventional tests. How much more efficiently is indicated in row 3 of the table, labelled "relative efficiency". These data, based on the Spearman-Brown equation, estimate for each test length how much the conventional tests would have to be lengthened to attain the reliability of the adaptive tests. For example, the adaptive test reliability at 5 items, .72 , is estimated to be equivalent to that of a conventional test 2.69 times as long, or 13.5 items in length. Notice that the relative efficiency of these adaptive tests always exceeds unity, but diminishes as test length increases. Thus, the adaptive tests are most advantageous, at least in terms of relative efficiency, at fairly short test lengths. At lengths of 10 or fewer items, these adaptive tests were at least 2.5 times as efficient as the conventional tests. But at lengths of 15 and more, the advantage, although still appreciable, is not quite so striking.

The advantage of adaptive tests is no so clear if we compare the validity of the two test types . Validity here is operationalized as the correlation between test scores and the examinee's raw score on the concurrently administered 50-item Word Knowledge test. From their superior reliability, one would expect the adaptive tests to be superior also in validity, at any constant test length. As Table 2 indicates, the validity advantage went to the adaptive tests at test lengths up to 10 items; at lengths of 15 and up, however, the conventional tests had slightly higher validity. I could find no explanation for this unexpected reversal. I must also note that none of the validity differences is statistically significant at the .05 level.

CONCLUSIONS

Based on the data reported above, several conclusions are offered with regard to the feasibility and psychometric merits of adaptive aptitude testing of Marine recruits.

1) Testing Marine recruits with CRT terminals is feasible, from both practical and human engineering standpoints. Embedded programmed instructions can effectively teach them the use of the testing terminals. The number ofproctors or attendants required to supervise and assist in the testing room appears to be acceptably small.

2) Striking psychometric efficiency was demonstrated for the adaptive tests of verbal ability used in this study. It appears that in military personnel testing applications, well-constructed short adaptive tests can achieve high levels of measurement reliability in less than half the number of items required using conventional testing procedures.

3) There is no apparent psychometric advantage to the intuitively appealing notion of variable-length adaptive tests, at least for the adaptive testing method used here.

4) Short, fixed-length adaptive tests of about 10 items per examinee seem to be sufficiently reliable for personnel testing purposes. The adaptive tests achieved a minimally satisfactory relaibility level (.80) in just 5 items; additional test length beyond 10 items did not yield psychometric returns proportional to the added administration time required.

REFERENCES

Bryson, R. A comparison of four methods of selecting items for computer-assisted testing. Technical Bulletin STB 72-8, San Diego: Naval Personnel and Training Research Laboratory, 1971.

Gorham, W.A. Opening remarks. In Gorham, W.A. (chair), *Computers and Testirg: Steps Toward the Inevitable Conquest*. Symposium presented at the 83rd Annual Convention of the American Psychological Association, Chicago, 1975.

Olivier, P. An evaluation of the self-scoring flexilevel testing model. Unpublished doctoral dissertation, Florida State University, 1974.

Owen, R.J. A Bayesian approach to tailored testing. Research Bulletin 69-92. Princeton, New Jersey: Educational Testing Service, 1969.

Owen, R.J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 1975, *70*, 351-356.

Samejima, F. Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, 1977,*42*, 193-198.

Seeley, L.C., Morton, M.A., and Anderson, A.A. Exploratory study of a sequential item test. Technical Research Note 129. Washington, D.C. U.S. Army Personnel Research Office, December, 1962.

Urry, V.W. A momte carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V.W. Tailored testing: a successful application of latent trait theory. *Journal of Educational Measurement*, 1977, *14*.

Waters, B.K. An empirical investigation of the stratified adaptive computerized testing model. *Applied Psychological Measurement*, 1977, *1*, 141-152.

Weiss, D.J. and Betz, N.E. Ability measurement: conventional or adaptive? Research Report 73-1, Psychometric Methods Program, Department of Psychology , University of Minnesota, 1973.

Weiss, D.J. Strategies of adaptive ability measurement. Research REport 74-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Weiss, D.J. Computerized adaptive ability measurement. Naval Research Reviews, November, 1975, 1-18.

Wood, R. Response-contingent testing. *Review of Educational REsearch*, 1973, *43*, 529-544.

Table 1. Testing time and examinee error summary for computer-administered test sessions. Each session consisted of programmed instruction, 60 experimental test items, and a 50-item criterion test.

	GROUP		
	A	C	
Test type	Adaptive	Conventional	Overall
Mean time (minutes)			
Total	70.5	60.7	
Instructions	9.5	10.3	9.0
Testing	61.0	50.4	
Errors			
Procedural errors	25	30	55
Proctor calls	5	12	17
Number of examinees	96	105	201

Table 2. Psychometric characteristics of the computer-administered verbal ability tests as a function of test type and test length.

Psychometric characteristic	Test type	Test Length					
		5	10	15	20	25	30
Reliability	adaptive (N = 96)	.79	.87	.88	.90	.91	.91
	conventional (N = 105)	.59	.73	.80	.83	.86	.89
Relative efficiency		2.7	2.5	1.9	1.8	1.7	1.3
Validity	adaptive (N = 93)	.77	.82	.83	.84	.85	.85
	conventional (N = 103)	.73	.81	.84	.85	.85	.87

Figure 1. The research design for administration of the experimental and criterion tests.

GROUP	TREATMENT (TESTS)				
	ADAPTIVE		CONVENTIONAL		CRITERION
	Form 1	Form 2	Form 1	Form 2	
A	X	X			X
C			X	X	X

Figure 2. Example examinee record (abbreviated).

	RAW SCORE		ABILITY ESTIMATE		POSTERIOR VARIANCE	
Form:	1	2	1	2	1	2
STAGE						
1	0	0	-.69	-.73	.548	.553
2	1	1	-.36	-.37	.401	.394
3	2	2	-.10	-.20	.332	.318
4	2	3	-.30	.02	.248	.266
5	3	4	-.14	.25	.229	.213
6	4	5	.01	.48	.193	.210
7	4	6	-.17	.65	.160	.184
8	5	6	-.05	.45	.145	.143
9	5	6	-.22	.26	.124	.115
10	6	7	-.15	..33	.115	.107
...	...		...		...	
30	20	21	.59	.97	.053	.048

Criterion score	27
Total time	57.3 minutes
Instruction time	8.5 "
Instruction errors	1
Proctor calls	0

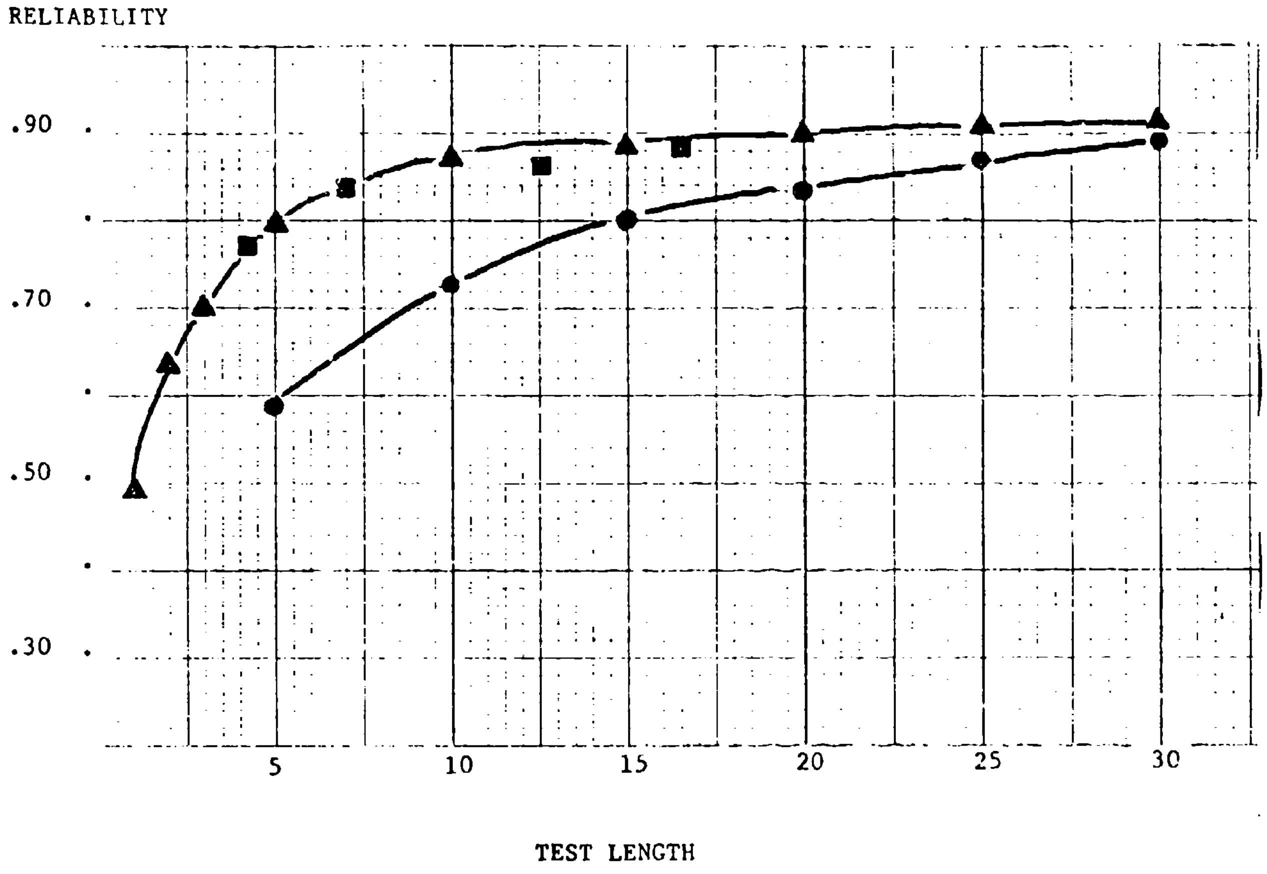


Figure 3. Alternate forms reliability plotted as a function of test length for the conventional an. adaptive tests.
Legend:

● conventional tesLs

▲ adaptive tests (fixed length)

■ adaptive tests (variable length)

THE EFFECTS OF ERRORS IN ESTIMATION OF ITEM CHARACTERISTIC CURVE PARAMETERS

Malcolm James Ree
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

The views expressed herein are those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

ABSTRACT

THE EFFECTS OF ERRORS IN ESTIMATION OF ITEM CHARACTERISTIC CURVE PARAMETERS

The Item Characteristics Curve (ICC) describes the probability of an individual answering a test question correctly as a function of the ability of that individual. It has been proven useful for estimating test scores, test reliability, normative equivalents, test equating, and adaptive testing purposes. The objective of this study was to determine the effects of errors in estimating the value of ICC parameters in the Birnbaum three parameter logistic model.

In a simulation, a standard set of ICC parameters was generated for 80 items by the normal random number generator. The standard values were generated to represent a set of typical five-option multiple choice items. These values were then systematically corrupted to reflect values obtained in operational item calibration procedures.

The known and corrupted parameters were compared and the results discussed. In general, the item location parameter *b* was the most influential in effect.

THE EFFECTS OF ERROKS IN ESTIMATION OF ITEM CHARACTERISTIC CURVE PARAMETERS

Malcolm James Ree

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

INTRODUCTION

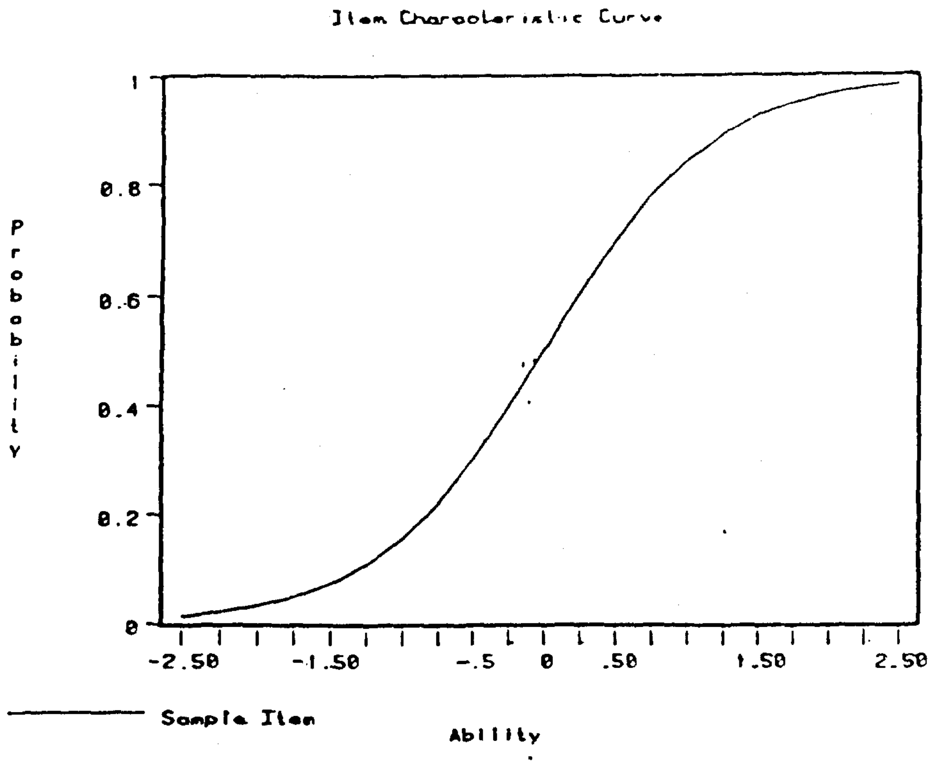
The requirement for precise descriptions of test items for use in computer driven adaptive testing, automated test item banking, and automated test construction has made the estimation of the Item Characteristice curve parameters important. This curve describes the probability of an individual answering a test question correctly as a function of the ability of that individual. It is useful for estimating test scores, test reliability, normative equivalents, test equating, and scoring responses during adaptive testing. The objective of this study is to determine the effects of errors in estimating the value of ICC parameters.

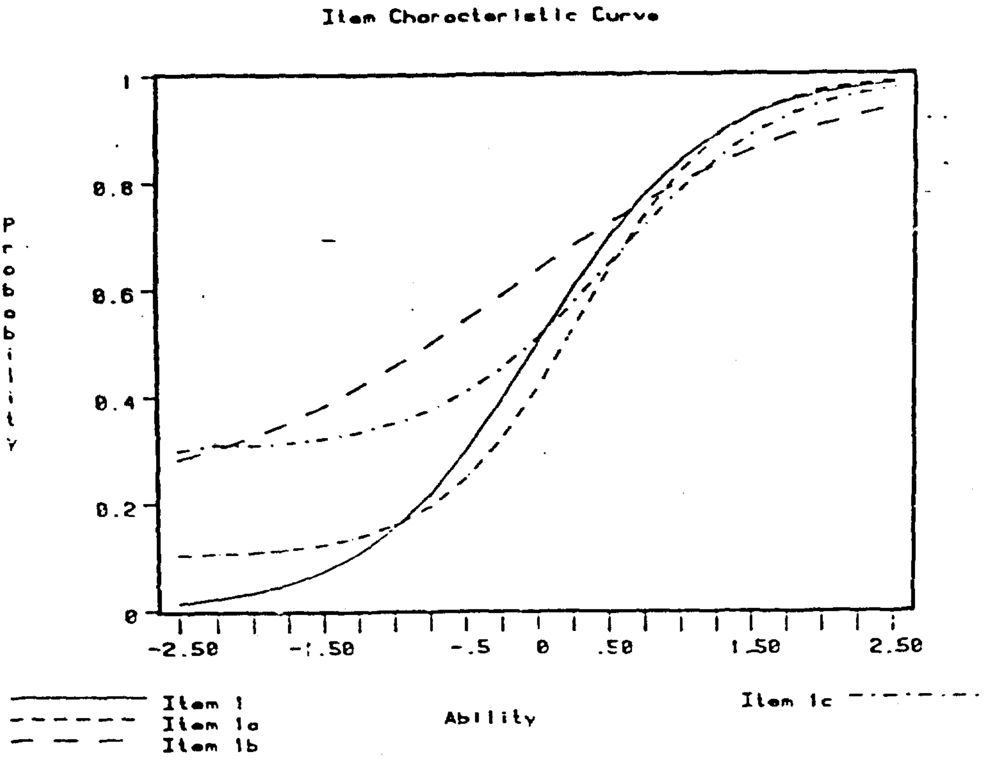
The Research Problem

The most frequently used model for estimating the ICC has been proposed by Birnbaum (1968). It is a logistic form which requires three parameters to describe the ICC. These are: *a*, item discrimination; *b*, item difficulty (or location); and *c*, the probability of chance success (or lower asymptote). The S shaped curve described by these parameters indicates the probability of answering a test item correctly as a function of ability. Figure 1 shows an ICC for a typical test item with $a = 1.0$, $b = 0.0$, and $c = 0.0$. The horizontal axis is scaled in terms of ability, which is denoted by theta (Θ). The probability of getting a correct answer at $\Theta = -0.5$, $\Theta = 0.0$, and $\Theta = 0.5$ may be found by reading the graph as .30, .50, .70, respectively. Figure 2 shows the same curve and the curves resulting from a misestimation of one of the ICC parameters. Note that the probability of correctly answering a test item varies from its true value as a function of errors in the estimation of ICC's. The questions to be answered in this study are: What is the effect of errors of estimation of ICC parameters? Is the effect of misestimation of any one parameter more serious than the misestimation of any other ICC parameter? And finally, does the misestimation of ICC parameters have a greater effect on estimation of ability in high or low ability subjects?

METHOD

A simulation was run in order to have known values against which to compare the effects of the errors of estimation. A standard set of ICC parameters was generated for 8C items by using the normal random number generator. The standard values were generated to represent a set of typical five-option multiple choice items. These values were then


Item Characteristic Curve
Probability
1
0.8
0.6
0.4
0.2
0
-2.50
-1.50
-.5
0
.50
1.50
2.50
Ability
Sample Item


Item Characteristic Curve
Probability
1
0.8
0.6
0.4
0.2
0
-2.50
-1.50
-.5
0
.50
1.50
2.50
Ability
Item 1
Item 1a
Item 1b
Item 1c

systematically corrupted to reflect values obtained by operational item calibration procedures observed in prior research (Ree, 1978). The erroneous parameters fall in the ranges in which errors might reasonably be expected to be found during actual estimation. The correlation between *b* and b_1, b_2, or b_3 is higher than between *a* and a_1, a_2, or a_3, which in turn is higher than between *c* and c_1, c_2, or c_3. This reflects the capability of operational procedures to produce accurate estimates of each ICC parameter. It is possible to estimate *b* better than *a*, and *c* is rather difficult to estimate under even the best of conditions.

In order to evaluate the effects of the misestimation of ICC parameters, two methods were used. Both were based on the relationship of "true score" (ξ) and "estimated true score" ($\hat{\xi}$).

True score is the sum of item scores which is given by:

$$P(\Theta)_j = c_i + (1 - c_i) \left(1 + e^{(-1.7a_i(\Theta - b_i))}\right)^{-1} \tag{1}$$

where $P(\Theta)_j$ is the probability of "subject" j answering the test item correctly and a_i, b_i, and c_i are the item parameters for item i.

Equation (2) defines true score.

$$\xi_j = \sum_{i=1}^{n} P(\Theta)_j \tag{2}$$

where ξ_j is the true score, n is the number of items, and $P(\Theta)_j$ is defined in equation (1). Similarly, the estimated true score is given by:

$$\hat{\xi} = \sum_{i=1}^{n} P(\hat{\Theta}) \tag{3}$$

where $P(\hat{\Theta})$ is computed from equation (1) using misestimated values of *a*, *b*, and *c*. The correlations of ξ and $\hat{\xi}$ and the average difference between ξ and $\hat{\xi}$ are good indicators of the effects of misestimation of ICC parameter because they indicate the linearity of the relationship of ξ and $\hat{\xi}$ and constant differences between ξ and $\hat{\xi}$.

Generating Errors of Estimation

Errors were estimated independently for each item parameter and for each degree of misestimation. The distribution of errors was specified to be normally distributed with a mean of zero and a variance designed to produce specified correlations between true ICC parameters and misestimated parameters.

In keeping with practical concerns and the available computer programs (see Ree, 1978), certain limits were placed on the misestimated parameters. These limits placed minimum and maximum values on the misestimated values.

The misestimated *a* parameters were limited to the range of 0.1 to +2.7, the misestimated *b* parameters from -2.9 to +2.9, and the misestimated *c* parameters to the interval between 0.00 and +0.30.

Simulation of Ability

Three distributions of ability, Θ, were generated using the random normal number generator. Distribution one was specified to represent the full range of ability, distribution two was specified to represent a group of high ability subjects, and the third distribution was specified to represent a group of low ability subjects. The shape of all three distributions was normal. The use of three differing distributions was necessary in order to detect if one or another ICC parameter played a larger role in specified regions of Θ

RESULTS AND DISCUSSION

Means, standard deviations, and correlations of the misestimated ICC parameters are presented in Table 1. The correlations (*r*) are representative of those achievable under actual item calibration procedures (Ree, 1978).

Table 2 presents the results of the true score analyses. High, Low, and Full Range Θ indicate the distribution of ability and $\mu\Theta$ and $\sigma\Theta$ indicate the mean and standard deviation of the ability group. The indices calculated in these analyses have practical significance for psychometric practice. If the correlation between ξ and $\hat{\xi}$ decreases due to misestimation of ICC parameters or if the average difference between ξ and $\hat{\xi}$ is large, errors of classification will ensue. These errors can be costly for both the examinee and for the agency doing the classification.

All of the ability groups were investigated with the same set of 80 test items. The items measure a wide range of ability along a single continuum, and this analytic procedure is equivalent to giving a moderately peaked, fixed length test to three selected groups of examinees.

Results of the correlational analysis indicate virtually no decrease in linearity of the relationship of ξ to $\hat{\xi}$ in any of the data sets. The observed correlations all range from .9898 to .9999, and it should be noted that the largest deviations were caused by misestimation of the *b* parameter in the low ability group. The percentile equivalent of the mean of this group is equal to about the 11th percentile in a normal population. Very few selection and classification decisions take place at this ability level.

In the Full Range sample, clearly the misestimation of the item *b* parameter causes the greatest problems. Examinees' abilities are, on the average, misestimated by 1.1 to 4.3 true score units. Misestimation of the *c* parameters is the next most influential, and misestimation of *a* being the least influential in this ability group.

Table 1. Description of True and Misestimated Item Characteristic Curve Parameters (N = 80)

Parameter	a	b	c
Mean	1.0141	.0773	.2057
S.D.	.3388	.9859	.0442
r	1.000	1.000	1.000
Parameter	a_1	b_1	c_1
Mean	.9762	.0197	.1970
S.D.	.3988	1.0449	.0682
r	.8178	.9079	.6321
Parameter	a_2	b_2	c_2
Mean	.9386	-.0379	.1801
S.D.	.5452	1.2642	.0960
r	.5427	.7211	.3850
Parameter	a_3	b_3	c_3
Mean	.9656	-.1670	.1587
S.D.	.7821	1.7400	.1208
r	.2789	.4348	.2153

Table 2. True Score, Estimated True Score, Difference Between True and Estimated True Score and Their Correlation Under Various Conditions of Misestimation of the ICC Parameters

ICC Parameters	$\Sigma P(\hat{\Theta})$	$\overline{DIF}$	r
	$\Sigma P(\Theta) = 45.5083$	Full Range Θ	
a_1 b c	45.7768	0.268	.9999
a_2 b c	46.2753	0.767	.9999
a_3 b c	46.4443	0.936	.9995
a b_1 c	46.6530	1.144	.9995
a b_2 c	47.7894	2.281	.9977
a b_3 c	49.6315	4.123	.9932
a b c_1	45.1093	-0.399	.9999
a b c_2	44.1093	-1.136	.9999
a b c_3	43.4196	-2.088	.9999
$\mu_\Theta = -0.06538$ $\sigma_\Theta = 1.04197$			
	$\Sigma P(\Theta) = 64.9861$	High Ability Θ	
a_1 b c	64.4318	-0.5543	.9999
a_2 b c	63.2667	-1.7194	.9999
a_3 b c	61.7106	-3.2755	.9998
a b_1 c	64.9483	-0.0378	.9999
a b_2 c	64.2062	-0.7799	.9996
a b_3 c	62.4761	-2.5100	.9979
a b c_1	64.8480	-0.1381	.9999
a b c_2	64.5581	-0.4280	.9999
a b c_3	64.1192	-0.8669	.9999
$\mu_\Theta = 1.17842$ $\sigma_\Theta = .34385$			
	$\Sigma P(\Theta) = 28.2354$	Low Ability Θ	
a_1 b c	29.3853	1.1499	.9999
a_2 b c	31.5541	3.3187	.9998
a_3 b c	33.5312	5.2958	.9998
a b_1 c	29.9667	1.7313	.9991
a b_2 c	32.7216	4.4862	.9955
a b_3 c	37.7748	9.5394	.9898
a b c_1	27.5781	-0.6573	.9999
a b c_2	26.4187	-1.8167	.9999
a b c_3	24.9947	-3.2407	.9999
$\mu_\Theta = -1.22157$ $\sigma_\Theta = .34385$			

The misestimation of the *a* parameter in the High Ability group is the most influential in distorting true score while extreme misestimation of *b* is also problematic.

In the Low Ability group the misestimation of the *b* parameter is of the greatest consequence, followed by *a* and then *c*.

It is worth noting that the ICC parameters are for items which are generally mismatched for the High and Low Ability groups, and if these parameters were made appropriate to the specified ability, as in adaptive testing procedures, the misestimation of the *b* parameter would be most influential in determining true score.

Results in the Low Ability group are the most surprising. Intuitively it would be expected that the *c* parameter should be most influential in distorting estimated true score. However, this was not found to be the case. It is wise to review the definition of *c* as the probability of getting an item correct at an infinitely low value of ability ($\Theta = -\infty$). An estimate of Θ of -2.9 is not infinitely low and most *c* parameters are measured at Θ locations which are considerably lower than -2.9 or even -4.0 which exceeds the limits specified for Θ in this study.

In a previous simulation study (Ree, 1978), the effects of misestimating all three simultaneously were determined. These effects are cumulative. The present study investigated the effects of misestimating each one while the other two remained perfect. In this manner the theoretical model may be explored piece by piece.

The item parameter that appears to be the most influential, *b*, is the item parameter which can be estimated with the most accuracy in practice while the *c* parameter, generally the least influential, is the item parameter which is the most difficult to estimate (see Ree, 1978).

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M.R., Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968 (Chapters 17-20).

Ree, M. Estimating item characteristic curves (AFHRL-TR-78-68). Personnel Research Division, Brooks AFB TX, 1978.

ANALYSIS OF AND COMMENTS ON DR. REE'S PAPER: "THE EFFECTS OF ERRORS IN ESTIMATION OF ITEM CHARACTERISTIC CURVE PARAMETERS"

Thomas A. Warm

U.S. Coast Guard Institute
P.O. Substation 18
Oklahoma City, Oklahoma 73169

The paper by Dr. Malcolm Ree is another of a series of creative studies he has been conducting on the practical problems of application of Item Response Theory. He is currently the most productive researcher in the critical area of parameter and ability estimation, and his conclusions are widely accepted as a basis for proceeding with other studies that build on his findings.

For that reason I took special care in studying his Monte Carlo experiment. With Dr. Ree's cooperation I obtained some of his supporting data, and set out to do some analysis of my own.

To my surprise, my conclusions were exactly opposite from his. I concluded that the c-parameter is most influential in estimating ability, the a-parameter is second most influential, and the b-parameter the least influential at typical levels of estimation error.

My difference with Dr. Ree stems from his statement that his corrupted "erroneous parameters fall in the ranges in which errors might reasonably be expected to be found during actual estimation" (page 4). I found that his three corruption levels (CL1, CL2, and CL3) are neither typical nor comparable. In fact, they are highly extreme and to differing degrees across parameters. That is, he inadvertently corrupted his a and b parameters two to five times as much as the c parameter. Thus, it is no wonder that he found the a and b parameters more influential than c.

To test the accuracy of his statement it is necessary to have a measure of parameter estimation error or corruption, and to have a standard or typical value of parameter estimation error for that measure.

I chose two different measures of corruption: (1) the correlation and (2) the root mean square error (RMSE) between the parameter (p) and its estimate ($\hat{p}$). The RMSE actually incorporates three different kinds of error as can be seen from its computational formula.

$$\text{RMSE}^2(p,\hat{p}) = \frac{\sum_{i=1}^{n}(p_i-\hat{p}_i)^2}{n} = (\bar{p}-\bar{\hat{p}})^2 + \sigma_p^2 + \sigma_{\hat{p}}^2 - 2r_{p,\hat{p}}\,\sigma_p\,\sigma_{\hat{p}}$$

where $\bar{p}$ and $\bar{\hat{p}}$, σ_p^2 and $\sigma_{\hat{p}}^2$, and $r_{p,\hat{p}}$ are the means, variances, and correlation of the parameter and its estimate, respectively. Thus, the RMSE is sensitive to errors of (1) the mean of the estimate, (2) the variance of the estimate, and (3) the correlation of the estimate with the parameter.

Let's see _if_ those types of errors were included in this study. Table I shows the mean ($\bar{\hat{p}}$), standard deviation ($SD_{\hat{p}}$), and correlations ($r_{p,\hat{p}}$) of the three parameters at each level of corruption in rows 1, 2, and 5, respectively. It is clear that all three types of errors were introduced by the corruption procedure.

To get an idea of whether the corruption levels of this study are "reasonable" or typical it is necessary to compare them to errors that have actually been found. Luckily, Dr. Ree (1978) has provided that information in a previous study (see Table II), part of which was conducted under almost identical conditions as the present study under discussion, except that no deliberate corruption was introduced. Both studies used the same sample size and distribution (of the full range group), the same test length, and very nearly the same parameters. The correlations and RMSEs in this previous study give us a good idea of what "might reasonably be expected to be found during actual estimation".

I also found another study (Gugel, et. al., 1976) which also gives comparison values. However, the Gugel (1976) study was conducted under considerably different conditions, and therefore, is not completely appropriate for comparison.

Figure I compares the correlations and RMSEs for each of the three parameters in this study (Ree, 1979) with those in Ree, 1978. I also included those from Gugel, 197C, just as a matter of interest.

From Figure I we can see that corruption level 1 (CL1) for the a-parameter contains slightly more error than the typical value from Ree, 1978. CL2 and CL3 contain amounts of corruption that are far from typical or reasonable to expect by both measures of corruption,r $(a,\hat{a})$ and RMSE $(a,\hat{a})$). For the b-parameter CL1 as well as CL2 and CL3 are grossly atypical for both measures. For the c-parameter it is an entirely different situation. CL1 and CL2 are much less corrupted than typically expected, while CL3 is about typical, by the r $(c,\hat{c})$ measure. By the RMSE $(c,\hat{c})$ measure CL1 is still grossly less than typical, whereas CL2 and CL3 are more than typical.

Therefore, we can conclude that each of the corruption levels are not in any sense equal across parameters. That is, CL1 does not contain equal corruption or error for each of the three parameters. Neither does CL2 nor CL3. Thus, it is not proper to make comparative judgments on the basis of the corruption levels.

To compare across parameters it is necessary to put the parameter corruptions on the same scale. I have done so by using the typical measures of error from Ree (1978) as a standard measure of corruption or typical error. My standard measure of corruption is the amount of corruption or error of a parameter at a corruption level divided by the comparable typical error from Ree (1978). To make a ratio of two numbers requires that they both be on a ratio scale of measurement. Since a correlation is not on a ratio scale, I converted the correlations to the standard error of estimate.

$$\text{SEE} = \sqrt{1 - r^2_{p,\hat{p}}}$$

I thus developed 2 standard measures of corruption, which are comparable between parameters.

$$\text{Std. SEE} = \frac{\sqrt{1 - r^2_{p,\hat{p}}}}{\sqrt{1 - r^2_{p,\hat{p}}}\ \boxed{\text{From Ree, 1978}}}$$

$$\text{Std. RMSE}\ (p,\hat{p}) = \frac{\text{RMSE}\ (p,\hat{p})}{\text{RMSE}\ (p,\hat{p})\ \boxed{\text{From Ree, 1978}}}$$

It is also necessary to consider the measure of the dependent variable, the error of true score (T). Ree's measure $(T - \hat{T})$ is the average error across all subjects. It is therefore only a measure of bias. I included in my dependent measures the correlation $r(T,\hat{T})$ and RMSE $(T,\hat{T})$.

I then plotted each of the dependent variable measures (of true score error) against each of the standard independent variable measures (of corruption) for each of the ability groups (i.e., full range, low and high). Incidentally, Figure II shows the distributions of ability for each of the three groups. Figures III through X show the results. In each figure I have listed my conclusion of the order of influence of the three parameters (most influential first).

Our area of interest is the amount of error on the dependent variable around the region of standard corruption = 1 on the independent variable. That level of corruption is typical or "what might reasonably be expected." A standard corruption = 2 or 3 means the corruption is 2 or 3 times what may reasonably be expected. Corruption 2 or 3 times what is reasonable or typical is not of interest to us except that it allows us to project the affect on the dependent variable at a reasonable level of corruption (i.e. = 1).

In Figure IIIA we see the affect of corruption on the absolute average error (not average absolute error) of true score for the full range group. At Std SEE = 1 we can see that the c-parameter causes the most error in estimated true score, followed by the b-parameter and then the a-parameter. This conclusion is the same for both Std SEE and Std RMSE (Figure IIIB) measures of corruption. In figure IV for the low ability group the c-parameter is again most influential with the a and b parameters about equal. In Figure VA for the high ability group the c-parameter is most influential, followed by the a-parameter and then the b-parameter. In Figure VB, using Std RMSE as a measure of corruption, the a and c parameters are most and about equally influential, followed by the b-parameter. However, since the slope of the c-parameter curve is much steeper at Std RMSE = 1, it is more influential than the a-parameter.

Figures VI, VII, and VIII show the same comparisons but with $[1 - r(T,\hat{T})]$ as the measure of true score error. In all cases there is no difference among the three parameters at a standard level of corruption. This result is not surprising, since the correlations differ only in the third or fourth decimal place. Thus, the three parameters do not differ in influence upon the correlation $r(T,\hat{T})$ at reasonable levels of parameter corruption.

Figures IX and X show the influence of parameter corruption on the RMSE of true score estimates (RMSE $(T,\hat{T})$) for the full range and low ability groups. RMSE $(T,\hat{T})$ for the high ability group was not available. In each case the c-parameter is most influential and usually with the a-parameter next most influential.

Table III is a summary of the influence of each of the parameters (two different measures), on each of the three measures of true score error for each of the three ability groups. Also included are the conclusions from the same comparisons, using the Gugel (1976) values as the standard (results not shown), which I also studied as a matter of interest.

In all 18 of the 18 cases in Table III where there is clearly a most influential parameter, that most influential parameter is the c-parameter. In 11 of the 15 cases where there is clearly a least influential parameter, that least influential parameter is the b-parameter. The b-parameter is never the most influential, and the c-parameter is never the least influential. In Figure I we saw that the Gugel (1976) values were considerably different from the Ree (1978) values. Nevertheless, the conclusions in Table III are very similar for both the Gugel (1976) standard and Ree (1978) standard. The fact that the general results are the same for two such varying sets of standards, indicates that what standard is used is not critical as long as it is somewhat plausible.

I, therefore, conclude that in general for the estimation of true score at typical levels of parameter corruption,

(1) the c-parameter is most influential,

(2) the a-parameter is next most influential, and

(3) the b-parameter is least influential.

Since the c-parameter is also the most poorly estimated parameter, our efforts should be concentrated on improving our estimates of it. We should also seek to improve our estimates of the a-parameter.

References

Gugel, J. F., Schmidt, F. L., and Urry, V. W. Effectiveness of the Ancillary Estimation Procedure. Proceedings of the First Conference on Computerized Adaptive Testing, Personnel Research and Development Center, U. S. Civil Service Commission, 1976, PS 75-6, 103-106.

Ree, M. J. Estimating Item Characteristic Curves. Personnel Research Division, Brooks Air Force Base, AFHRL-TR-78-68, Nov. 1978.

Ree, M. J. The Effects of Errors in Estimation of Item Characteristic Curve Parameters. Proceedings of the 21st Annual Conference of the Military Testing Association, 1979.

Warm, T. A. A Primer of Item Response Theory. U. S. Coast Guard Institute, Tech. Rept. 940278, Feb. 1978.

		$\hat{a}$				$\hat{b}$				$\hat{c}$		
p	a	CL1	CL2	CL3	b	CL1	CL2	CL3	c	CL1	CL2	CL3
$\bar{p}$	1.014	.976	.939	.906	.077	.020	-.038	-.167	.206	.197	.180	.159
SDp	.339	.399	.545	.782	.986	1.045	1.264	1.740	.044	.068	.096	.121
Minimum		.10	.10	.10		-2.9	-2.9	-2.9		.00	.00	.00
Maximum		2.7	2.7	2.7		2.9	2.9	2.9		.30	.30	.30
r(p,$\hat{p}$)		.818	.543	.279		.908	.721	.435		.632	.385	.215
SEE		.575	.840	.960		.419	.693	.900		.775	.923	.977
RMSE(p,$\hat{p}$)		.233	.467	.768		.444	.886	1.602		.054	.093	.128
Std. SEE		1.052	1.535	1.756		1.922	3.179	4.130		.796	.948	1.003
Std. RMSE		1.474	2.953	4.825		1.989	3.975	7.184		.640	1.105	1.523
FULL RANGE GROUP												
$\hat{T}$-T		.268	.767	.936		1.144	2.281	4.123		-.399	-1.136	-2.088
r(T,$\hat{T}$)		.9999	.9999	.9995		.9995	.9977	.9932		.9999	.9999	.9999
RMSE(T,$\hat{T}$)		.733	2.134	3.565		1.381	3.100	6.205		.495	1.280	2.304
LOW ABILITY GROUP												
$\hat{T}$-T		1.150	3.319	5.296		1.731	4.486	9.539		-.657	-1.817	-3.241
r(T,$\hat{T}$)		.9999	.9998	.9998		.9991	.9955	.9898		.9999	.9999	.9999
RMSE(T,$\hat{T}$)		1.187	3.411	5.444		1.748	4.614	9.751		.668	1.832	3.263
HIGH ABILITY GROUP												
$\hat{T}$-T		-.554	-1.719	-3.276		-.038	-.780	-2.510		-.138	-.428	-.867
r(T,$\hat{T}$)		.9999	.9999	.9998		.9999	.9996	.9979		.9999	.9999	.9999
RMSE(T,$\hat{T}$)		N/A	N/A	N/A		N/A	N/A	N/A		N/A	N/A	N/A

Table I. Descriptive measures of parameters and true scores for each of 3 corruption levels, and each of 3 ability groups (N/A=not available).

					Range				Error		
p	$\bar{p}$	σ_p	r(p,$\hat{p}$)	RMSE	Min.	Max.	Skew.	Kurt.	$\bar{e}_p$	σe_p	r(p,e_p)
a	.950	.284			.46	1.6					
$\hat{a}$	1.035	.288	.837	.158	.56	2.0			.085	.148	-.290
b	.164	.929			-1.65	1.97					
$\hat{b}$	.196	1.000	.976	.223	-1.80	2.26			.032	.221	.213
c	.200	.046			.087	.348					
$\hat{c}$	.188	.082	.225	.0837	.059	.376			-.012	.083	-.320
T	45.326	14.204			0	80					
$\hat{T}$	45.158	14.044	.9999	.306	0	80			-.168	.256	-.631
θ	-.0127	1.0191			-3.84	3.67	-.0050	3.1144			
$\hat{\theta}$	.0706	.9899	.965	.280	-2.50	2.50			.0833	.267	-.239

Table II. Parameter statistics from Ree, 1978.

Standard Corruption Measure		Ree,1978						Gugel, 1976					
		Standard SEE			Standard RMSE			Standard SEE			Standard RMSE		
Group		Full Range	Low Group	High Group	Full Range	Low Group	High Group	Full Range	Low Group	High Group	Full Range	Low Group	High Group
True Score Error Measure	$\hat{T}-T$	c b a	c a=b	c a b	c b a	c a=b	a=c b	c a b	c a b	c a b	c b a	c b a	a=c b
	$r_{T,\hat{T}}$	a=b=c	a=b=c	a=b=c	a=b=c	a=b=c	a=b=c	a=b=c	a=b=c	a=b=c	a=b=c	a=b=c	a=b=c
	RMSE $T,\hat{T}$	c a=b	c a=b	N/A	c a=b	c a b	N/A	c a b	c a b	N/A	c a b	c a b	N/A

Table III. Parameters listed in order of influence (high, medium, low) on 3 measures of true score error for 2 measures of standard corruption with 2 standards for each of 3 ability groups.

	$\hat{a}$				$\hat{b}$				$\hat{c}$			
	Ree78	CL1	CL2	CL3	Ree78	CL1	CL2	CL3	Ree78	CL1	CL2	CL3
$\bar{e}_p$	.085	-.038	-.075	-.108	.032	-.057	-.115	-.244	-.012	-.009	-.026	-.047
σ_{e_p}	.148	.230	.461	.760	.221	.440	.879	1.583	.083	.053	.089	.119
r_{p,e_p}	-.290	-.055	-.093	-.159	.213	-.084	-.085	-.145	-.320	-.019	-.079	-.151
FULL RANGE GROUP												
$\bar{e}_T$	-.168	.268	.767	.936		1.144	2.281	4.123		-.399	-1.136	-2.088
σ_{e_T}	.256	.682	1.991	3.440		.774	2.099	4.637		.293	.590	.974
r_{T,e_T}	-.631	-.956	-.996	-.995		-.814	-.909	-.966		.692	.930	.972
LOW ABILITY GROUP												
$\bar{e}_T$		1.150	3.319	5.296		1.731	4.486	9.539		-.657	-1.817	-3.241
σ_{e_T}		.294	.787	1.261		.243	1.079	2.022		.121	.234	.378
r_{T,e_T}		-.965	-.991	-.999		-.020	-.909	-.963		.730	.951	.971

Table IV. Descriptive statistics of typical errors (Ree,78), of corruption level errors of parameters introduced by the corruption procedure, and of the true score error caused by the corrupted parameters. Typical true score error statistics (Ree,78) are in the 1st column of the Full Range Group.

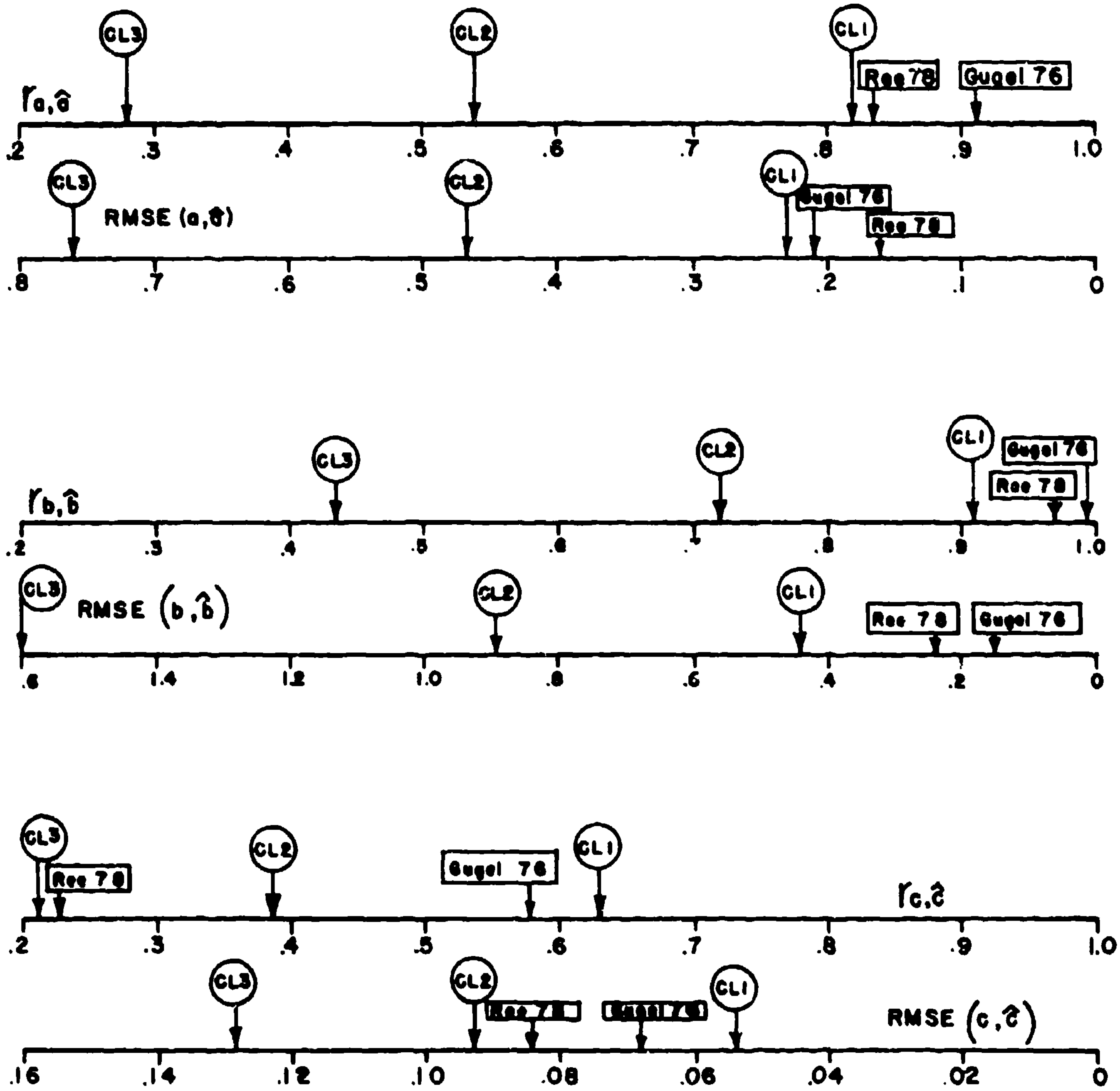


FIGURE I. Comparison of $r\ (P,\hat{P})$ and RMSE $(P,\hat{P})$ of parameter corruption levels (CL) and standard levels of corruption Ree (1978), and Gugel, et al (1976).

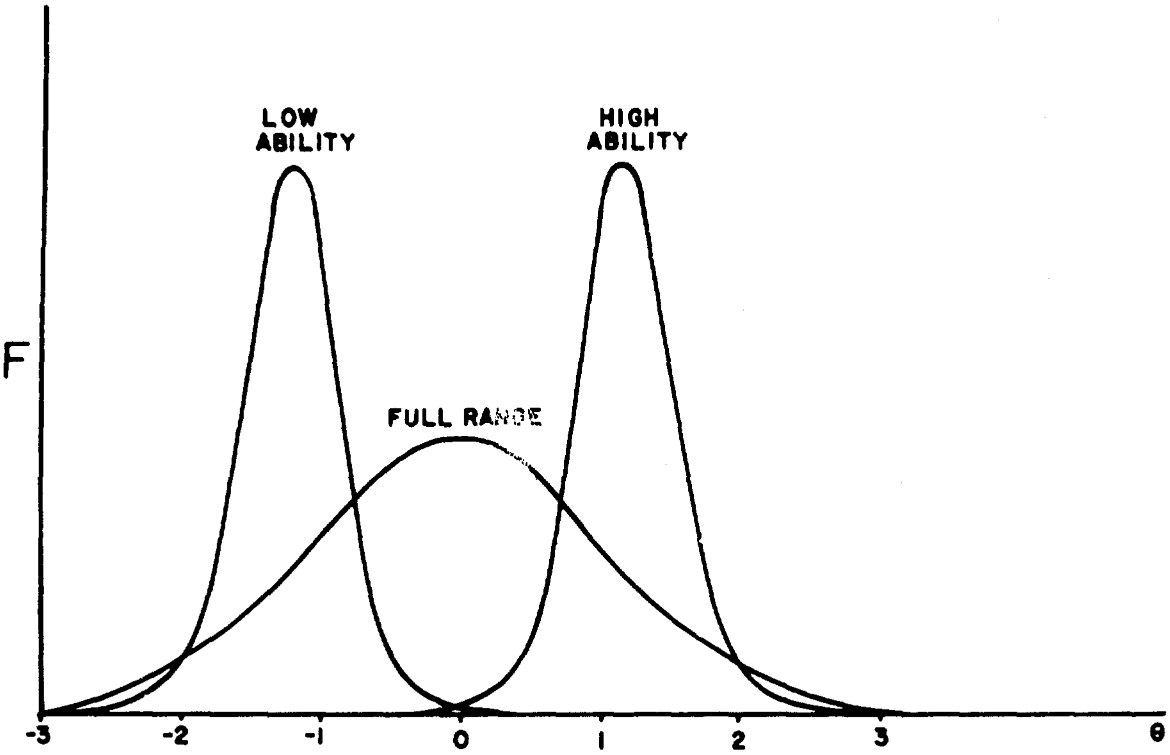


FIGURE II DISTRIBUTION OF ABILITY OF FULL RANGE, LOW ABILITY AND HIGH ABILITY GROUPS.

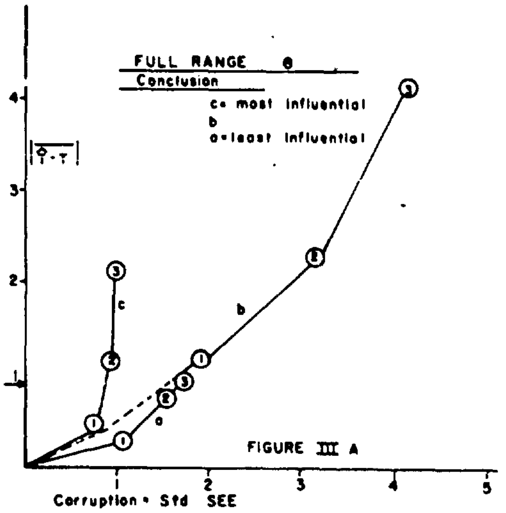


FIGURE III A

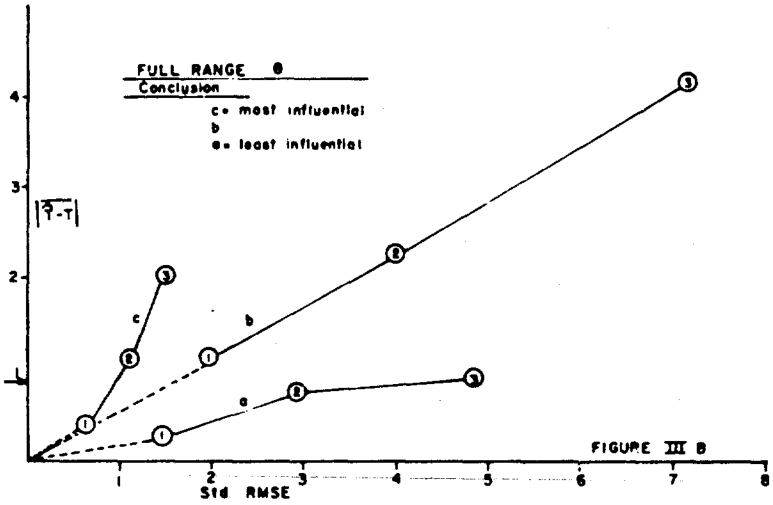


FIGURE III B

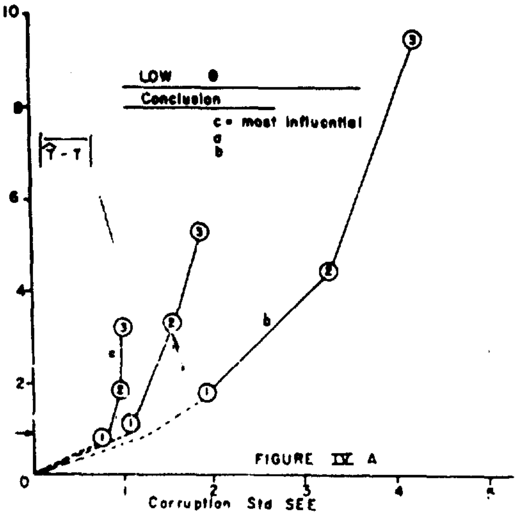


FIGURE IV A

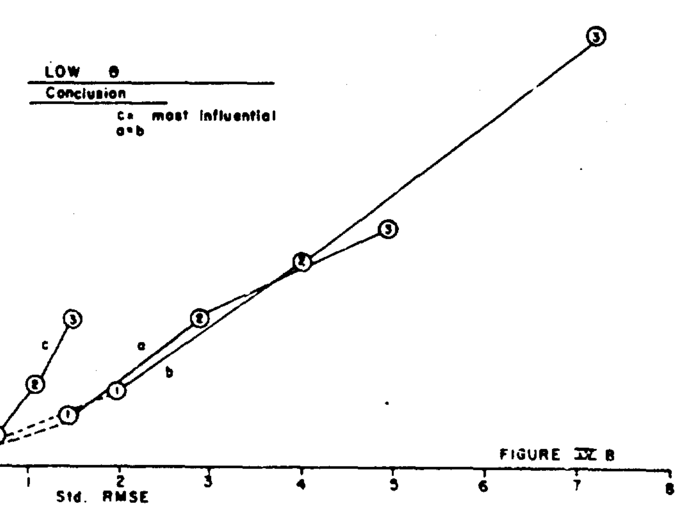


FIGURE IV B

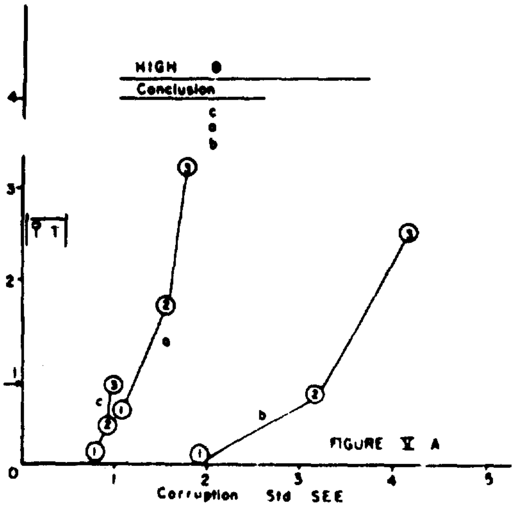


FIGURE V A

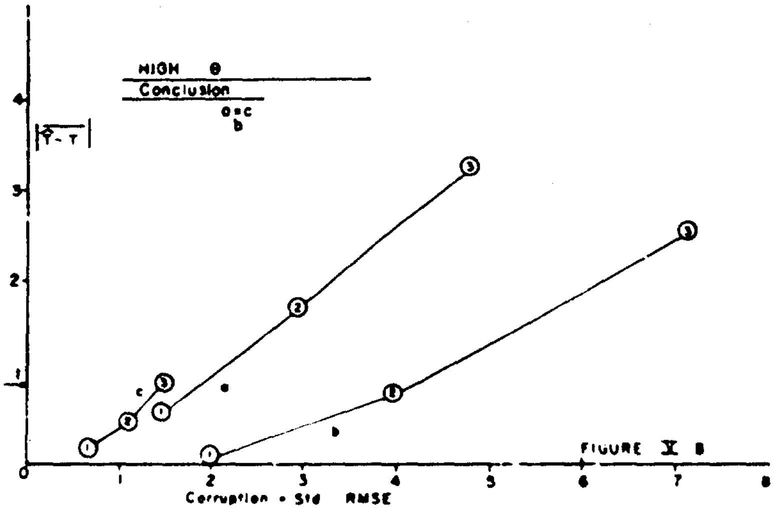


FIGURE V B

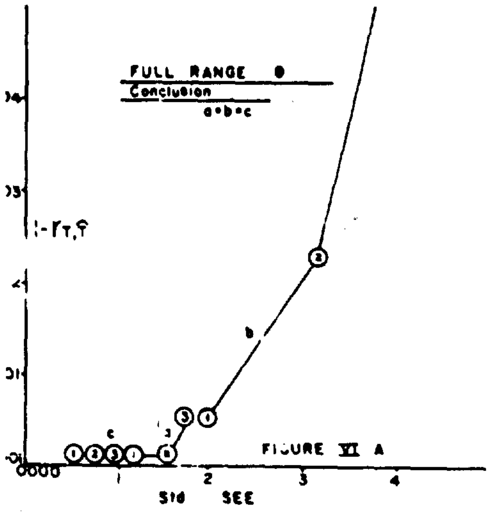


FIGURE VI A

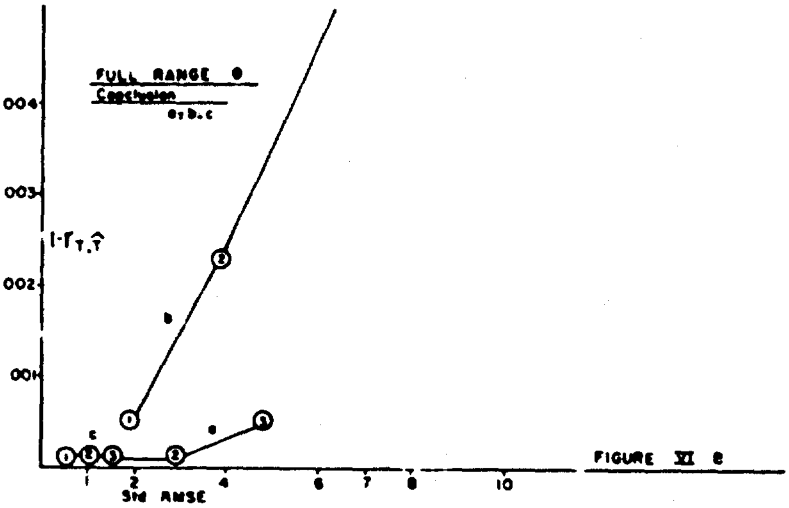


FIGURE VI B

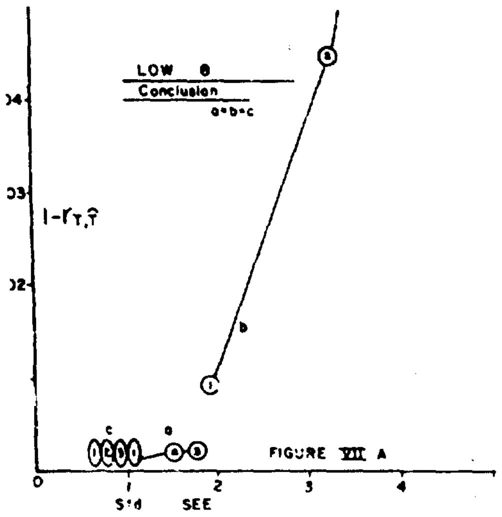


FIGURE VII A

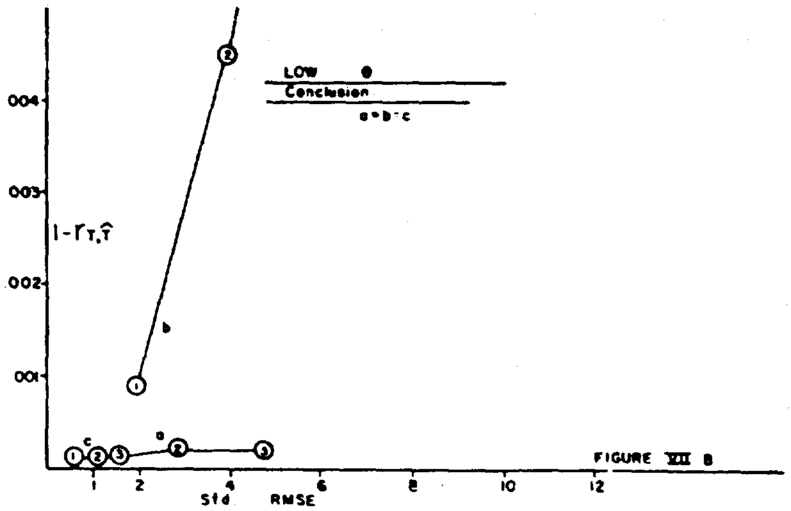


FIGURE VII B

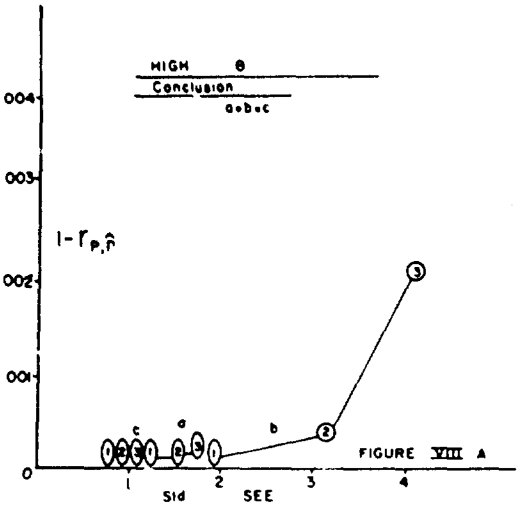


FIGURE VIII A

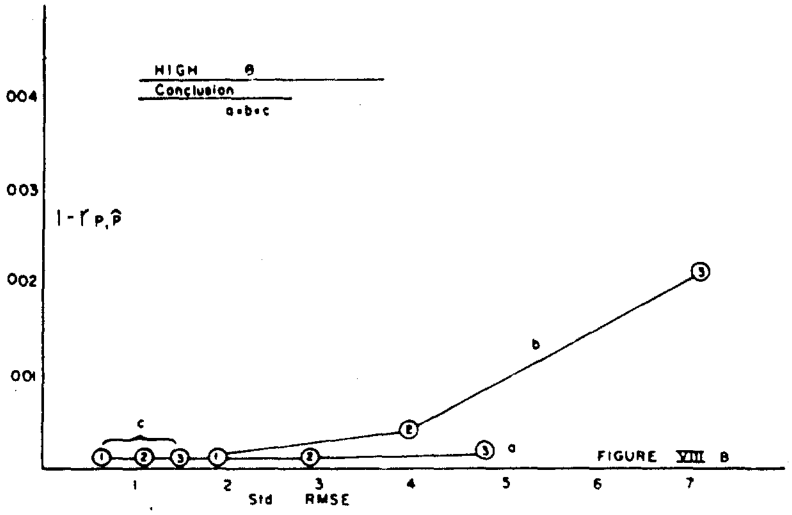


FIGURE VIII B

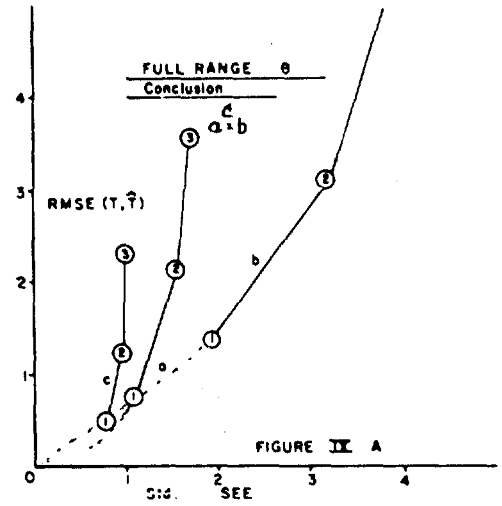


FIGURE IX A

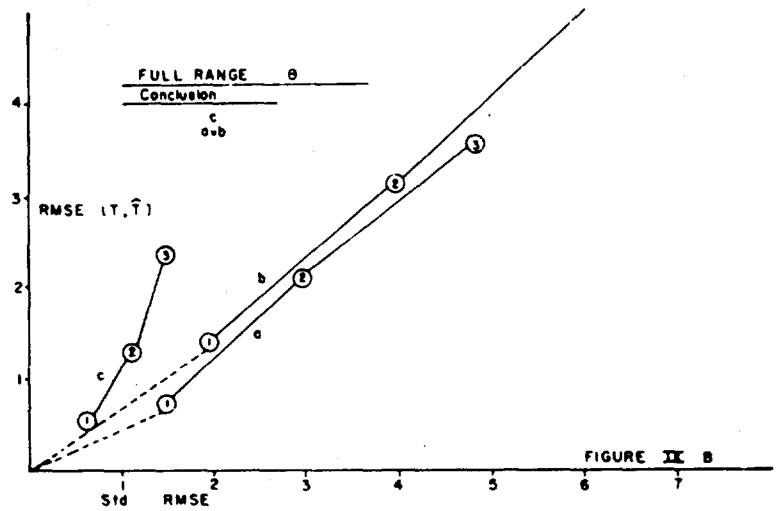


FIGURE IX B

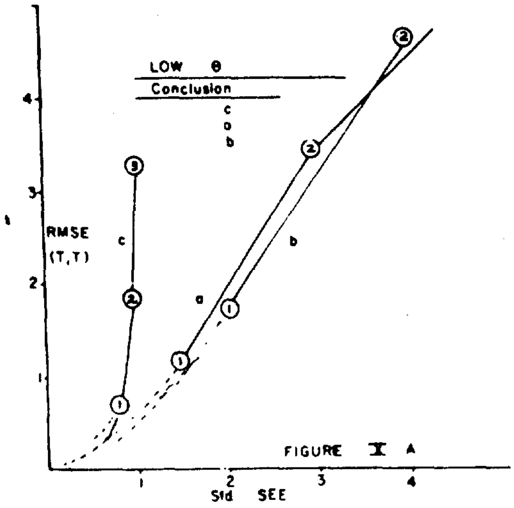


FIGURE X A

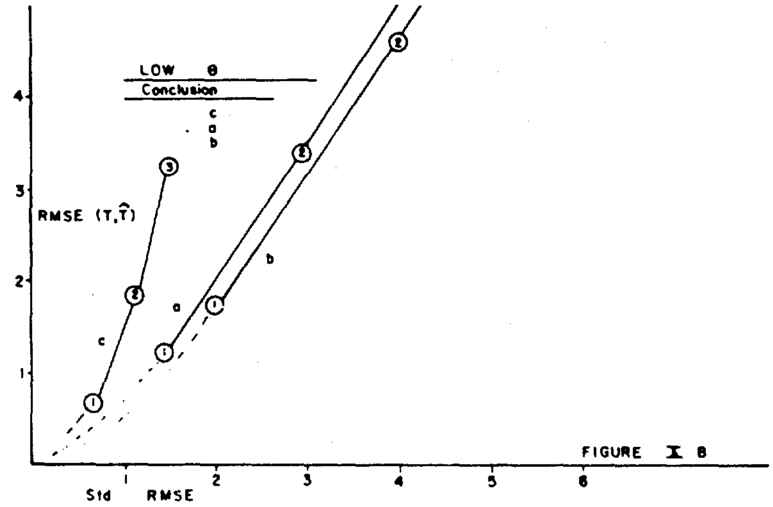


FIGURE X B

AUTOMATED PERFORMANCE TESTING IN NAVY TECHNICAL TRAINING

Marc Hamovitch

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

The Navy operates a large scale computer managed instruction (CMI) system which trains over 9,000 students on a daily basis. One of the goals of this system is to train students to particular performance objectives in the shortest possible period of time. Although knowledge testing in CMI courses is automated, most performance skills are still manually tested and scored. Very often with these manual procedures, testing sessions are lengthy, scoring is inaccurate, and no records of errors are kept. This lack of detailed information on past performance errors hampers remediation efforts, and further prolongs training time.

These delays were particularly noticeable in the Radioman (RM) "A" School. As part of the requirements of the RM course, a student is required to learn how to use a Teletypewriter (TTY). While the device itself is electric, the testing procedures and scoring were manual. These procedures resulted in lengthy testing sessions and poor remediation techniques. To reduce these delays, we at the Navy Personnel Research and Development Center (NPRDC) developed an Automated Performance Testing (APT) program to time and score tests automatically. The program provides detailed error information that can be used to guide more effective practice and remediation.

The primary objective of this research was to find out if there were, in fact, procedures for providing students with this detailed information on past performance errors that would lead to reduced overall training time. This required the development of the most appropriate techniques for presenting and distributing the feedback to teletyping students in the school. Another objective was to determine the effect of automatically testing performance on the length of testing sessions.

Previous research in the area of APT has focused primarily on methods of instrucion rather than on scoring methods and remediation techniques (Dierks, 1977; Dixon, 1976; Krag & Van Brunt, 1970; Pask, 1958; Peterson & Staples, 1969; Rainey, 1970; Sharkey & Thomas, 1973; Sherrill, 1976; Showel, 1974; Wolcott, 1976). Noticeably missing from previous systems is any automated recording of specific typing errors. Most systems simply note that an error has been made. The student is supposed to keep typing while trying to reduce the number of errors and increase speed. With the system developed for the RM School, specific errors were recorded and stored for cumulation and eventual feedback to the student. We felt that by informing the student as to what his mistakes were, he could concentrate on just those errors and not have to slow down to concentrate on everything.

Before I discuss what we did in the study, I'd like to give a brief description of the procedures in the RM School. The course of instruction consists of individualized instruction in common core knowledge and TTY training followed by either a sea- or shore-oriented practical deck watch. The common core phase includes instruction in basic Navy message forms and handling procedures, and is conducted in conjunction with daily TTY training sessions. These two phases are coordinated by interspersing TTY practice and test sessions between completion of instructional modules. Depending upon how fast a student achieves knowledge goals, he could be assigned from zero to three TTY assignments a day. Each session is scheduled to last two hours. Certain intermediate levels of TTY skill are necessary before other portions of the course can be attempted. Thus, some students are held at the TTY assignment beyond the normal two-hour session until they reach the specified level of performance. To complete the TTY training, the student must be able to type 700 keystrokes (functions) in a five minute period with no more than five keystroke errors. This is roughly equivalent to 28 words per minute of conventional typing.

The CMI system directs a student through TTY training with procedures that are independent of the method of testing. Students are given up to 9 sessions to build their speed up to 300 function per 5 minutes. If they have not already done so, students are then required to take a test to qualify at 300 functions. The student may take up to three tests to qualify. If the student fails all three, he is assigned to mandatory night study. This process is repeated for each level at which the student must qualify (300, 400, 500, 600, and 700 functions). The number of practice sessions allowed varies with the qualification level. Because students cannot progress through certain parts of the course until they have qualified at 500 functions, and other parts until 700 functions, a restriction is placed on the students at these levels. Until the student qualifies, CMI keeps assigning TTY tests instead of interspersing instructional modules. Continued failure leads to increasingly severe action, ultimately involving the academic review board and possible removal from the course.

When these procedures were implemented manually, during a pilot study of the CMI version of the course, all phases of TTY instruction took place in a room with 25 TTY's. Testing was conducted at one of five times during the day. All three tests were taken during the same testing session by all students who were in the room and needed to take a test. The students waited for their names to be called, got their test booklets and when everyone was ready, the instructor started a timer. After each test the student marked on his paper how many functions he had typed. After all three tests, the student went over each test and counted the errors. Then the instructor checked the tests of those students who felt they had qualified. The instructor then filled out an administrative computer form which the student fed into an optical scanner. Naturally there were delays at each point in the process causing much lost time. As it was, due to the nature of the tests, an instructor could never be sure a student hadn't made errors that didn't show up on the paper that was turned in.

Under the APT system, students took tests and practiced in a room of 86 TTY's. The 30 machines in the front of the room were used exclusively for testing as they were connected to the scoring computer. Students could take a test whenever they wanted to rather than only at one of five specified times. The student received a test from the instructor, then entered administrative information at the TTY. The next key the student hit started his five minute test. Then, the keyboard locked up and began typing the student's results which were already passed on to the CMI computer. A copy of the entire testing session is shown in Figure 1. Under these procedures, if the student passed on the first or second test of a series, all three would not have to be taken. Also, if a student failed the first test, he could go back and practice before taking another.

EXPERIMENT 1

Method

The first study that was conducted involved a comparison of the manual and automated testing procedures on the length of the testing session. Ten testing sessions were observed under the manual procedures and ten under automated condidions. The duration of four basic groups of activities were measured. These included recording course information, testing, scoring, and obtaining the next CMI assignment.

Results and Discussion

The results, shown in Table 1, indicate that the major difference between manual and automated testing procedures is in scoring time. Under the manual system, scoring time increases as the number of students taking tests increases, while under APT, scoring is virtually instantaneous regardless of the number of students being tested. The one minute duration for scoring is primarily due to the time to print out the results. Zero is listed for the time to get the next assignment under APT as this is virtualy complete by the time the results have finished printing. Even under the longest APT situation (3 tests) the session takes less time than the shortest manual session (1 student) indicating the substantial improvement that APT represents.

EXPERIMENT 2

The second experiment was conducted to evaluate a by-product of APT called the Error Distribution Report (EDR). We designed this computer-generated report to take advantage of the error data collected by the APT program. The report, shown in Figure 2 includes information on which key was hit, and which should have been hit. It was designed to provide students and instructors with a precise description of the students' errors and where they occurred. Thus, practice could be more effective thereby reducing training time.

```
T I M E S  STARTED      0820781412
ENTER YOUR LAST NAME                              HAMO

NOW ENTER YOUR S-S-N  IE 999-99-9999     563-74-5706

NOW ENTER YOUR COURSE NUMBER -3- DIGITS 040

NOW ENTER YOUR ITEM NUMBER -6- DIGITS   041704

NOW ENTER YOUR TEST NUMBER -2- DIGITS O 040

TEST NUMBER INVALID -- REENTER  .  12

YOU ENTERED THE FOLLOWING - - - -
NAME:       HAMO
SSN:        563745706
COURSE:     040
ITEM:       04'704
TEST:       12
IF CORRECT TYPE -YES- IF NOT -NO-       Y.

YOUR TIME BEGINS WHEN YOU KEY THE FIRST FUNCTION



NSS DE NBRG 001/27
PTTCZYUW RUHHLHH0457 3322330-CCCC--RUWJBGF RUECORR RUWJIAA RUWNDUA.
ZNY CCCCC
P 272320Z NOV 75
FM COMNAVBASE PEARL HARBOR HI
TO RUWJBGF/USS KING
RUECORR/USS BAUSELL
INFO RUWJIAA/COMDESRON FIFTEEN
RUWNDUA/COMCRUDESPAC SAN DIEGO CA
BT
C O N F I D E N T I A L DRILL //N01462//
DEPLOYMENT OF OPERATING FORCES (U)
A. COMCRUDESPAC LTR 1520.1A.
1. (C).SVCS ARRANGED TO ACCOMMODATE KING AND BAUSELL FOR UPKEEP.
2. (C) CRAIG PRESENTLY SKED FOR FUEL SERVICES ON 20 DEC.
ADS UPON ARR SUBIC
BT
/0457



NNNN
OLTC DE NEEX 004/25
PTTCZYUW RUWJEEX0045 3321456-CCCC--RUHHLHA RUWNDUA RUWJAJA.
ZNY CCCCC


TEST OVER, PLEASE WAIT

STUDENT NAME  HAMO                    SSN  563745708
 ITEM #  041704  TEST #  12       COURSE #040
YOU HAVE TYPED   598 FUNCTIONS WHICH QUALIFIES YOU
AT THE   500 FUNCTION LEVEL
YOU MADE THE FOLLOWING ERRORS:
MSG FORMAT #    ERRORS DESCRIPTION
 1  12 LINE 05    1     SUBSTITUTION
 X  XX LINE XX    2     GARBAGE LINE
 X  XX LINE XX    2     FATAL GARBAGE LINE
 X  XX LINE XX    2     FATAL GARBAGE LINE
 X  XX LINE XX    2     FATAL GARBAGE LINE
 X  XX LINE XX    2     FATAL GARBAGE LINE
ADDITIONAL LINES WITH ERRORS NOT DETAILED
TOTAL KEY FUNCTIONS TYPED  599
ESCESSIVE ERRORS. GRADING TERMINATED PREMATURELY.

YOUR NEXT ASSIGNMENT WILL BE PRINTED AT YOUR
LEARNING CENTER CLUSTER.  TEAR OFF THI~ SUMMARY
STATEMENT AND REPORT TO YOUR LEARNING
SUPERVISOR.               103143
```

TABLE 1

Mean Number of Minutes Spent at Test Related Activity
for Manual and Automated TTY Testing

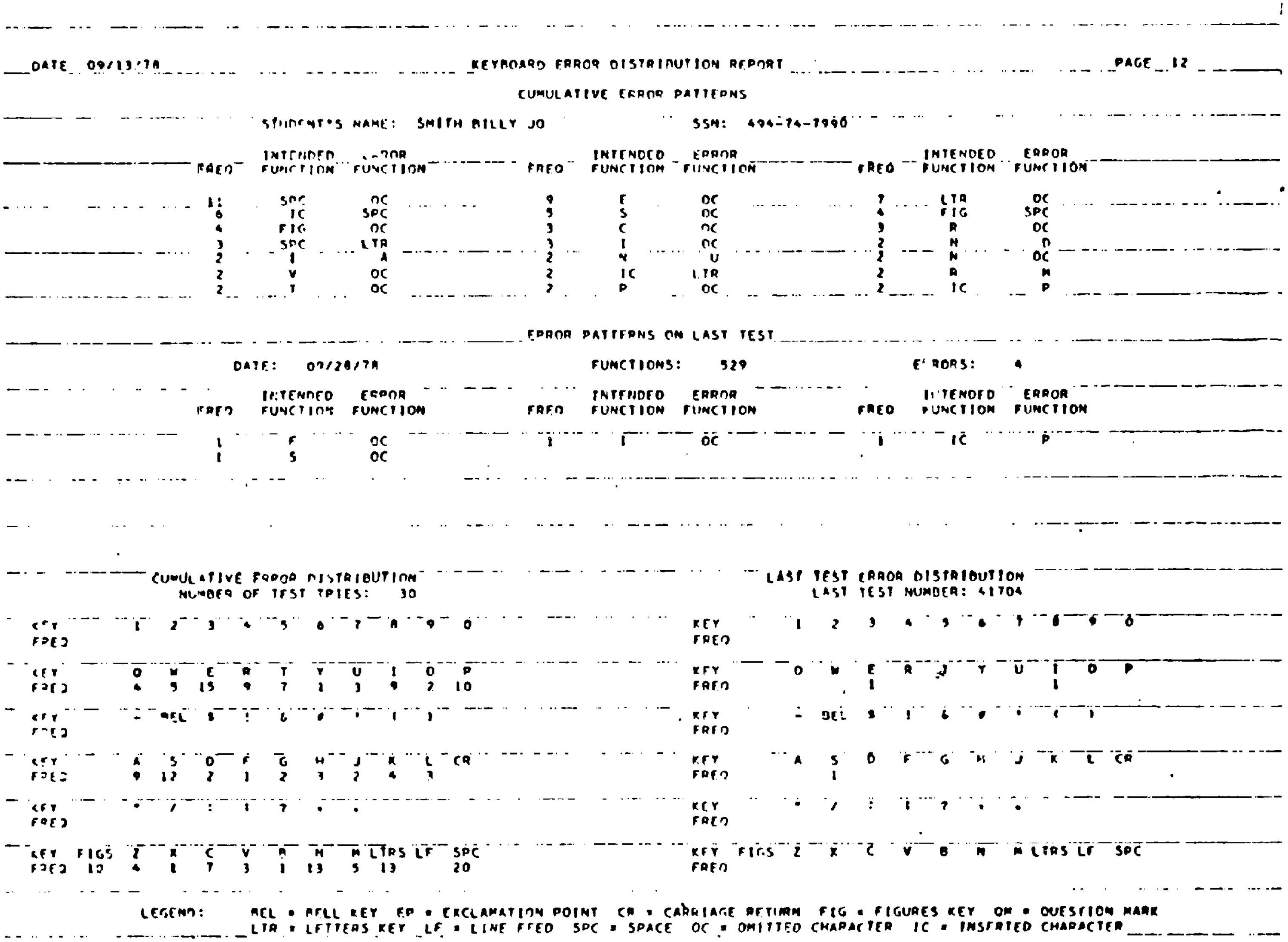
	TTY Test Condition	
Activity	Manual	Automated
Course Information	2	2
Testing[a]	20	5
Scoring[b]		
1 student	1	1
15 students	8	
30 students	15	
Next Assignment	3	0

	Number of Students			Number of Tests		
	1	15	30	1	2	3
Totals	26	33	40	8	16	24

Note. Data rounded to nearest minute.

[a] Based on minimum number of tests (i.e., 3 for manual testing, 1 for automated testing).

[b] Based on .5 minutes per student for manual group.



```
DATE  09/13/78                        KEYBOARD ERROR DISTRIBUTION REPORT                                PAGE  12
                                            CUMULATIVE ERROR PATTERNS
                      STUDENT'S NAME:  SMITH BILLY JO               SSN:  494-74-7990

                 INTENDED   ERROR                   INTENDED   ERROR                   INTENDED   ERROR
           FREQ  FUNCTION   FUNCTION          FREQ  FUNCTION   FUNCTION          FREQ  FUNCTION   FUNCTION

            11     SPC        OC                9      E          OC               7     LTR        OC
             6     IC         SPC               5      S          OC               4     FIG        SPC
             4     FIG        OC                3      C          OC               3     R          OC
             3     SPC        LTR               3      I          OC               2     N          D
             2     I          A                 2      N          U                2     N          OC
             2     V          OC                2      IC         LTR              2     R          M
             2     T          OC                2      P          OC               2     IC         P

                                           ERROR PATTERNS ON LAST TEST

                   DATE:   09/28/78                 FUNCTIONS:    529                ERRORS:    4
                 INTENDED   ERROR                   INTENDED   ERROR                   INTENDED   ERROR
           FREQ  FUNCTION   FUNCTION          FREQ  FUNCTION   FUNCTION          FREQ  FUNCTION   FUNCTION

             1     F          OC                1      I          OC               1     IC         P
             1     S          OC


           CUMULATIVE ERROR DISTRIBUTION                              LAST TEST ERROR DISTRIBUTION
             NUMBER OF TEST TRIES:   30                                  LAST TEST NUMBER: 41704

  KEY          1   2   3   4   5   6   7   8   9   0                KEY          1   2   3   4   5   6   7   8   9   0
  FREQ                                                              FREQ

  KEY          Q   W   E   R   T   Y   U   I   O   P                KEY          Q   W   E   R   T   Y   U   I   O   P
  FREQ         4   5  15   9   7   1   3   9   2  10                FREQ                 1                   1

  KEY          -  BEL  $   !   &   #   '   (   )                    KEY          -  DEL  $   !   &   #   '   (   )
  FREQ                                                              FREQ

  KEY          A   S   D   F   G   H   J   K   L  CR                KEY          A   S   D   F   G   H   J   K   L  CR
  FREQ         9  12   2   1   2   3   2   4   3                    FREQ             1

  KEY          "   /   :   ;   ?   ,   .                            KEY          "   /   :   ;   ?   ,   .
  FREQ                                                              FREQ

  KEY  FIGS    Z   X   C   V   B   N   M LTRS LF  SPC               KEY  FIGS    Z   X   C   V   B   N   M LTRS LF  SPC
  FREQ  10     4   1   7   3   1  13   5  13      20                FREQ

           LEGEND:     BEL = BELL KEY  EP = EXCLAMATION POINT  CR = CARRIAGE RETURN  FIG = FIGURES KEY  QM = QUESTION MARK
                       LTR = LETTERS KEY  LF = LINE FEED  SPC = SPACE  OC = OMITTED CHARACTER  IC = INSERTED CHARACTER
```

Figure 2. Error Distribution Report

Method

Three groups of twenty students differed on how frequently they were given reports. Students in a Daily (D) group received reports every day, those in a Remedial (R) group received reoprts only when they were assigned to night study, and students in a Never (N) group never received reports. The experimenter explained the reports the first couple of times they were given to the students, and simply handed them out thereafter.

Results and Discussion

The results revealed a differential attrition rate among the groups. Twenty-five percent of the Remedial and Never groups dropped out for typing related reasons. Only ten percent of the Daily group attrited for those reasons. This fifteen percent reduction in attrition translates into the avoidance of substantial losses of resources that would be wasted on students who attrite. The results also revealed faster completion times of the TTY portion of the course for the groups that received reports. The data, adjusted for the differential attrition rates, are shown in Table 2. About two training days were eliminated from the typing completion times of the group that never received reports. With these positive results, the RM School has fully implemented the error distribution reports. It should be noted that APT is not restricted to use in an already computerized system. This particular program that scores typing tests can be modified to score almost any kind of typing related activity. Thus, it can be seen that automated performance testing can be of substantial value in a military training environment by reducing both training time and attrition.

TABLE 2

Mean Number of Hours to Criteria

		TTY Qualification Criteria (Functions)					
GROUP		T1[a]	300	400	500	600	700
D	$\bar{X}$	14.9	18.2	22.7	32.3	52.2	68.3
	SD	(7.1)	(8.9)	(10.4)	(13.3)	(28.3)	(32.4)
R	$\bar{X}$	13.6	19.7	24.1	35.4	53.7	68.7
	SD	(6.0)	(8.0)	(10.2)	(14.4)	(32.8)	(35.1)
N	$\bar{X}$	17.3	24.6	30.3	44.8	64.8	82.4
	SD	(5.7)	(8.5)	(9.9)	(13.9)	(38.7)	(36.4)

Note. Data adjusted for differential attrition. N's are 13, 12, 12 for Groups D, N, and R, respectively.

[a] Point at which first test was taken.

REFERENCES

Dierks, C. J. Evaluation of an automated touch typing system, Colorado Journal of Educational Research, 1977, 17, 1-7, 14.

Dixon, S. Jr. The effect of letter-level technique on the reduction of errors in intermediate high school typewriting (Doctoral dissertation, Michigan State University, 1976). Dissertation Abstracts International, 1976, 37, 3357A-3358A. (University Microfilms No. 76-27, 038).

Dvorak, A., Merrick, N. I., Dealey, W. L., & Ford, G. C. Typewriting Behavior. American Book Co., 1936.

Krag, E., & Van Brunt, R. E. Training clerical help, Training and Development Journal, 1970, , 36-39.

Nathanson, Y. S. A "Conceptual" Basis of Habit Modification, Journal of Applied Psychology. 1929, 13, 469-485.

Pask, G. Electronic keyboard teaching machines, in A. A. Tumsdaine and R. Glaser (Eds.) Teaching Machines and Programmed Learning--A Source Book. Department of Audio-Visual Instruction: National Education Association, 1960.

Peterson, J. C., & Staples, J. Declare war on undetected typing errors, Business Edcuation World, 1969, 49, 9-24.

Rainey, C. M. Comparison of instructional methods upon psychomotor performance of students varying in dexterity in beginning typewriting (Doctoral dissertation, University of Missouri-Columbia, 1976). Dissertation Abstracts International. 1977, 37, 5781A. (University Microfilms No. 77-4942).

Sharkey, V. J., & Thomas, R. C. The evaluation of two automated systems for teaching typewriting (NTEC Tech. Rep. IH-220). Orlando, FL: Naval Training Equipment Center, March 1973.

Sherrill, J. L. Comparison of three typing training methods (Doctoral Dissertation, Indiana University, 1975). Dissertation Abstracts International, 1976, 36, 6047A-6048A. (University Microfilms No. 76-6347).

Showel, M. A comparison of alternative media for teaching beginning typists, The Journal of Educational Research, 1974, 67, 279-285.

Wolcott, J. M. The effect of computer-assisted instruction, traditional instruction, and locus of control on achievement of beginning typewriting students (Doctoral dissertation, Temple University, 1976). Dissertation Abstracts International, 1976, 37, 1942A-1943A. (University Microfilms No. 76-22, 070).

COST-EFFECTIVENESS OF COMPUTER-BASED INSTRUCTION FOR MILITARY TRAINING*

Jesse Orlansky
Institute for Defense Analysis

and

Marty Rockway
Air Force Human Resources Laboratory

INTRODUCTION

The military services have supported much of the R&D which led to the many current applications of computer based instruction in both the military and civilian sectors. As a result of these efforts there is little doubt that computers, when properly used, can be effective tools for the delivery and management of instruction. However, despite more than two decades of computer applications to instruction, the issue of whether or not the use of the computer is cost-effective when compared to the other alternatives for performing the same functions was still open to question. This paper describes the results of a recent review (Orlansky and String, 1979) of computer-based instruction in military training. The primary purpose of this review was to determine if the available data is sufficient to permit any definitive conclusions concerning the cost-effectiveness of CBI in military training environments.

METHODS OF INSTRUCTION

The Rand Corporation's "Method of Designing Instructional Alternatives (MODIA)" identifies 20 different methods of teaching (Carpenter-Huffmen, 1977). For convenience in this report the major methods of instruction are grouped into the four general categories of "Conventional", "Individualized", "Computer-Assisted" and "Computer-Managed" instruction.

Conventional Instruction. Conventional instruction refers to many possible combinations of lectures, discussions, laboratory, and tutorial sessions as a method of instruction. A key feature of conventional instruction is that groups of students proceed through a course at the same pace. Differences in the amount of information retained by students are reflected in their grades at the end of the course.

* Adapted from a study of the same title by Orlansky and String for presentation at the 1979 Conference of the Military Testing Association in San Diego, CA on 16 Oct 1979.

Individualized Instruction. In individualized instruction, a course is arranged in a series of lessons and tests and each student proceeds at his own pace. Mastery of each lesson is prescribed as a condition of progress. Differences among students are reflected in how long it takes them to complete a course, although grades may also be given. All methods of computer-based instruction rely on some form of individualized instruction; by definition, the term "individualized instruction" will be used here to apply only to this method of instruction conducted without computer support.

Computer-Assisted Instruction (CAI). In this paper the term computer-based instruction refers generally to both CAI and CMI methods of instruction. In computer-assisted instruction (CAI), the student interacts in real time, via an interactive terminal, with instructional material that is stored in the computer. Most CAI systems diagnose student performance, prescribe lessons, and maintain student records. Among the major systems used primarily in a CAI mode for military instruction are PLATO (Programmed Logic for Automatic Teaching Operations), TICCIT (Time Shared Interactive Computer-Controlled Television), LTS (Lincoln Terminal System), and GETS (General Electric Training System).

Computer-Managed Instruction (CMI). In computer-managed instruction (CMI), instruction takes place away from the computer. The computer scores tests and interprets results to the student; advises on following or alternative lessons; recommends remediation; and manages student records, resources, and administrative data. This process is initiated typically when the student places a test answer sheet on an optical reader connected to the central computer. He receives the results on a printout which tells him how well he performed, what lesson to take next, and where to find it. In most CMI systems CAI may be included as another instructional option. Among the major systems used primarily in a CMI mode for military instruction are the Air Force Advanced Instructional System (AIS) at Lowry AFB, the Navy Computer Managed Instructional System at NAS Memphis, and the Army Computerized Training System (CTS) at Ft Gordon.

DISTINCTIONS BETWEEN MILITARY AND NON-MILITARY INSTRUCTION

Military personnel receive pay and allowances while they are in training. Thus, any procedure which can reduce the length of time required for training, without significantly affecting the amount and/or quality of information acquired, can assist in reducing the cost of training at military schools; it can also result in increasing the amount of time spent by military personnel in operational assignments during their military careers. Military training courses are designed to qualify students for well-defined jobs to which they can be assigned upon successful completion of these courses.

The situation differs in almost all types of public and private education where students remain at school for required periods of time and are not paid while being instructed. These schools receive no direct benefits for completing instruction in less than the required time. Courses are generally not designed to qualify students for particular jobs and, obviously, schools cannot assign students to jobs when they graduate.

A major consequence of these distinctions is that methods of instruction that are cost-effective for military training may not be cost-effective in other areas. Another is that research on computer-based instruction supported by the military Services has emphasized the possibility of saving student time while maintaining student achievement constant. Research on instruction in non-military settings has been concerned more with the amount of student achievement at the completion of a course than with the amount of time needed by students to acquire the material.

MAJOR FINDINGS AND DISCUSSION

The use of computer-assisted and computer-managed instruction in military training has been evaluated in about 30 studies (producing 48 sets of data) since 1968. Most (70 percent) of the data on CAI come from experiments with few students (up to 50) and limited course materials (1 day to 1 week). There are fewer studies of CMI but these involve more students (600 to 2500) and longer courses (2 to 10 months). There is a wide range of subject matter in these studies, e.g., knowledge, theory, and hands-on performance skills; electronics machinist, recipe conversion, vehicle repair, fire-control technician.

Each of the 30 studies report effectiveness. However, only eight of the studies which report effectiveness also provide some cost data. The latter data are limited to expenses incurred during the experiment and are incomplete with respect to costs of program management, maintenance and repair, instructional support, and other factors important in determining life-cycle costs. It is probably inappropriate to extrapolate from cost data in experiments to the costs of large-scale, long-term operational training programs.

The comparisons of alternative methods of instruction are limited. Generally, CAI or CMI is compared to conventional instruction. There are only a few comparisons of CAI and CMI with individualized instruction (without computer support), a comparison which relates to the benefits of computer support. In addition, time savings found when CAI or CMI are compared to conventional instruction may be due to a combination of self-pacing, computer support, revised and possibly reduced amounts of course materials.

In the remainder of this section some of the major findings will be summerized and discussed.

EFFECTIVENESS OF CAI AND CMI

Based on evidence provided by military research studies and qualified as noted above, the effectiveness of CAI and CMI is evaluated as follows:

Student achievement. Student achievement at school with CAI is about the same as that with conventional instruction in most comparisons and superior in about one-third of the comparisons. The differences in achievement are not thought to have practical significance. Student achievement with CMI is about the same as that with conventional instruction. These findings are important but also inevitable because students are held in CAI and CMI courses until they achieve at least the standards established previously for conventional instruction.

Student time savings. Students instructed by CAI or CMI save about 30 percent (median value) of the time required to complete the same courses given by conventional instruction. There is a wide range in amounts of time reported as saved in these studies. The amounts of time saved by CAI and CMI cannot be compared directly because different courses were used for tests of these methods of instruction. Where courses have been given for relatively long times, the initial student time savings are maintained and, despite monthly fluctuations, tend to increase. This finding is based on four courses given by the Air Force Advanced Instructional System for about 4 years and on three courses given by the Navy Computer Managed Instruction System for about 15 months; both systems are CMI systems.

Student attrition. The academic elimination rates in four courses on the Air Force Advanced Instructional System (AIS) appear to have increased slightly over 4 years compared to the previous base rates; however, the average academic elimination rate for all courses at Lowry AFB, i.e., those not on AIS, increased at the same time. Similar increases in attrition seem to have occurred in six courses on the Navy CMI system over a 15-month period; attrition dropped in oen course; data on non-CMI course for the same time period were not provided.

Attitudes of students and instructors. Students generally prefer CAI or CMI to conventional instruction. The attitudes of instructors are reported only in a few studies but these are almost always unfavorable to CAI and CMI in comparison to conventional instruction.

Time savings found with individualized instruction and computer-based instruction. Some data were found where the same course was given bv conventional instruction, individualized instruction (i.e., self-paced instruction, individualized instruction without computer support) and either CAI or CMI. Individualized instruction saves student time. However, the addition of computer support (either CAI or CMI) to individualized instruction does not increase the amount of student time saved very much beyond that achieved by individualized instruction alone (i.e., without computer support). Again, differences between time savings attributed to CAI and CMI cannot be evaluated because different courses were used in each group of studies. These data do not necessarily imply that the addition of CAI or CMI to individualized instruction (i.e., transforming the method of instruction) is not cost-effective. That would depend on whether the incremental costs of computer support are offset by benefits in other areas such as, e.g., a need for fewer instructors and support personnel and ior less administrative support.

COSTS OF INSTRUCTION IN MILITARY TRAINING

The benefits of computer-based instruction have to be compared with the cost providing this type of instruction, but only incomplete cost data were found.

Collection of cost data. The military Services maintain systems that report the costs of individual courses. These are useful for such purposes as settin, reimbursement rates for training students from other Services or other governments. They are not useful for analyses of the costs of different methods of instruction for the following reasons: (1) they do not distinguish the costs of parts of a course, which would permit determining the costs of different methods of instruction used within a course; (2) costs of training support and management, that may vary considerably between methods of instruction, are allocated to individual courses on essentially arbitrary bases, such as the student load of all courses.

Type of data needed on cost of instruction. Each method of instruction in military training requires the expenditure of funds for most, but not necessarily all, of the following functions:

PROGRAM DEVELOPMENT

- Program Design
- Instructional Materials
  - Conventional Instiuction
  - Individualized Instruction
    - Programming
    - First-Unit Production
  - Computer-Based Instruction
    - Programming
    - Coding

PROGRAM DELIVERY

Instruction

Instructors
Instructional Support Personnel

Equipment and Services

Laboratory (including simulators)
Media Devices
Computer Systems
Communications

Materials (including Consumables)

Facilities

PROGRAM MANAGEMENT AND ADMINISTRATION

STUDENT PERSONNEL

Pay and Allowances

Other (Permanent Change of Station, Temporary Duty)

Limited cost data were found for some of these resources. However, cost data were not found or were extremely limited for the following resources for all methods of instruction:

- Program Design
- Instructional Material: conventional instruction
- Instructional Support Personnel
- Laboratory Equipment
- Materials (including consumables
- Program Management and Administration
- Student Personnel: Permanent Change of Station, Temporary Duty, etc.

Collection of More Complete Data. Detailed cost data, required for analytical purposes, may be collected in three possible ways:

- Universal, more complete reporting for all courses and support functions.
- Sampling selected courses and support functions
- Ad hoc

The costs and benefits of these ways of collecting the cost data needed to evaluate alternative methods of instruction should be examined.

COST-EFFECTIVENESS OF CAI AND CMI

There have been only a few attempts to estimate the cost-effectiveness of CAI and CMI and these are based on incomplete analyses of the costs of instruction. All of them are based on the premise that the amount of student training time saved by a method of instruction provides major cost savings; the amounts of cost savings are estimated by computing the pay and allowances of students for the amounts of student time saved in training; the resultant amounts should more properly be called "cost avoidance savings". Four of these studies consider other costs in addition to those avoided by student time savings, such as for preparing course materials, purchase or use of computers, and the number of instructors required by each method of instruction.

The dollar amounts of such "savings" could be large, depending, of course, on the number of students assumed for these estimates, e.g., about $10 million a year for about 50,000 students instructed in FY 1977 by the Navy CMI system and about $3 million a year for about 3500 students instructed in FY 1978 by the Air Force AIS system. According to two cost-effectiveness evaluations that have been reported, the PLATO IV CAI system is judged to be not as cost-effective as individualized instruction. These conclusions are based on incomplete cost data in two small-scale tests. The Air Force AIS was found to be cost-effective, compared to instructor supported, self-paced instruction in one course (Inventory Management) but not in three others; the computer costs which made the latter courses not cost-effective were judged to be small in comparison to other school costs (AIS Service Test, 1978). Since all of these findings are based on incomplete cost data, the findings cannot be generalized or even taken seriously.

Other benefits, beyond those of saving student training time, are often said to occur with CAI and CMI, largely because the computer can compile records and direct the attention of instructors, on the basis of various algorithms. The following list is illustrative rather than complete:

- More precise data for improving and updating course materials
- Improved control over equipment, facilities, and materials for instruction
- Improved allocation of resources among students
- Improved ability to accommodate fluctuations in student loads
- Increased student: instructor ratios, as well as the ability to use some instructors with less advanced qualifications
- Reduced need for support by noninstructional personnel
- Reduced time of students on base waiting for courses to start
- Reduced time of students on base waiting for orders after completing courses
- Improved integration of records of students at school with those in central, computer-based personnel files
- Improved utilization of instructors.

Many of these benefits may occur with the use of CAI and CMI. None of them have been included in any cost-effectiveness evaluation known to us. Records kept at Lowry AFB for students instructed by the AIS show that, compared to prior periods, they spend less time waiting to enter a course and waiting for an assignment after completing a course. Records kept by the Navy CMI system show that the average on-board count of students in school has been reduced for those instructed by that system: the extent to which this may be attributed to various benefits has not been examined.

SUMMARY

The potential value of computer-assisted and computer-managed instruction for military training rests primarily on findings that (1) computer-assisted and computer-managed instruction save 30 percent or more of the time (median value) required by students under conventional instruction and that (2) student achievement at school is about the same with computer-assisted and computer-managed instruction as with conventional instruction. However, these results do not necessarily imply that computer-assisted and computer-managed instruction are cost-effective because of fundamental problems with the measures of effectiveness and of cost used in the studies from which these results are taken. Effectiveness, as measured by student achievement at school, is not necessarily a measure of performance by course graduates in relevant jobs after they leave school. Data on the costs of alternative methods of instruction reported in various studies are essentially incomplete, particularly with respect to courseware, student; instructor ratios, support and management services; this applies both to computer-based and conventional instruction. The results that have been reported are limited to obvious costs observed during experiments (e.g., preparation of courseware, rental of computers) and do not consider long-term costs associated with operational applications (e.g., numbers of instructors and support personnel, revisions to course materials, maintenance of software and facilities, management).

CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

1. The effectiveness of computer-assisted and computer-managed instruction for military training has been measured only by student achievement at school and not by performance on the job. Correlations between performance at school and on the job have not been established for any method of instruction.

2. Student achievement in courses at military training schools with computer-assisted instruction is the same as or greater than that with conventional instruction. Student achievement in courses with computer-managed instruction is about the same as that with conventional instruction.

3. Computer-assisted and computer-managed instruction in military training save about 30 percent of the time (median value) needed by students to complete the same courses given by conventional instruction. The amounts of time reported as saved vary widely, but little attention has been given to the factors that could account for the wide variation. Most of the results on computer-assisted instruction come from experiments of limited duration, with limited amounts of course materials, and with relatively few students. Where computer-managed instruction has been used for extended periods (up to 4 years), the initial time savings have been maintained or increased.

4. Individualized instruction (self-paced instruction without computer-support) saves student time; little or no additional student time is saved when the same courses are given by computer-assisted or computer-managed instruction.

5. Attitudes of students toward computer-assisted and computer-managed instruction appear to be favorable. Attitudes of instructors are reported as unfavorable, but this finding is based on very limited data. In addition, little attention has been given to the role of instructors in computer-based instruction and to how they should be prepared for this type of instruction.

6. Only limited and incomplete data are available on the costs of computer-assisted and computer-managed instruction in military training. Data that are collected routinely on the costs of operational training programs are too highly aggregated, particularly with respect to training support functions, for use in analytical comparisons of computer-based instruction with conventional instruction.

7. Estimates based on the amounts of student time saved suggest that the Navy Computer Managed Instruction System avoided costs of about $10 million in FY 1977 and that the Air Force Advanced Instructional System avoided costs of about $3 million in FY 1978. These estimates are incomplete because they do not consider the other costs of providing computer-managed instruction at these installations or compare these costs with the costs of alternative methods of instruction for the same courses.

RECOMMENDATIONS

1. Improve methods currently available for measuring performance on the job in areas related to technical training. Compare achievement at school with performance on the job for students in courses given by computer-assisted and computer-managed instruction; to whatever extent opportunities exist, do the same thing for the same courses given by conventional and individualized instruction. The job-performance data should be collected for several time intervals after students leave school to determine whether benefits in favor of any method of instruction are sustained as job experience increases.

2. Evaluate alternative methods of collecting reliable data on the costs and effectiveness of instruction in military training. Based on these findings, develop and initiate data-collection programs on the costs and effectiveness of alternative methods of instruction.

3. Bring up to date the "Integrated Department of Defense Plan for Research and Development on Computers in Education and Training" (Department of Defense, September 1975). Support is needed for Exploratory and Advanced Development (6.2 and 6.3 RDT&E funds) on many subjects identified in this paper, such as the development of objective measures of performance on the job, comparisons of student achievement at school with performance on the job, the development of methods to measure the quality of course materials and delivery of instruction, and studies to account for the relative contributions of self-pacing, course revision, computer support, and other factors to the amounts of student time saved by computer-assisted and computer-managed instruction. Support for other studies to improve various aspects of computer-assisted and computer-managed instruction may well be questioned until more reliable cost data are available to determine areas of high pay-off.

4. Collect data on the costs of instruction for courses and course segments given now by computer-assisted or computer-managed instruction for military training, e.g., PLATO IV at Sheppard Air Force Base, Texas, and at Chanute Air Force Base, Illinois; TICCIT at North Island Naval Air Station, San Diego, California; Advanced Instructional System at Lowry Air Force Base, Denver, Colorado; and Navy Computer Managed Instruction System at Naval Air Technical Training Center, Millington, Tennessee. Comparable baseline cost data should also be collected, as far as possible, for alternative methods of instruction for the same courses. Projections of cost should be made for computer-managed instruction systems that are now being planned; i.e., the Navy Aviation Training Support System, the Army Automated Instructional Management System, and the Marine Corps Communication-Electronics School CAI/CMI System.

5. Determine the factors which account for the large variations in the amounts of student time saved by computer-assisted and computer-managed instruction in various studies. Consideration should be given to such factors as quality of courseware (including that in conventional courses), instructional strategy, types of subject matter presented in courses, and the amount and type of guidance provided by instructors. An effort should also be made to resolve the extent to which such factors as self-pacing, course revision, shortening courses, and various types of computer-support contribute to the total amounts of student time saved.

6. Determine the extent to which observed increases of student attrition with computer-managed instruction are due to this method of instruction and to other factors that may also be present, such as changes in the quality of students.

7. Determine the attitudes of instructors to computer-based and other methods of instruction in a systematic manner so that remedial actions can be taken as required.

MEASURING TRAINING EFFECTIVENESS:
PROCEDURAL VERSUS COMBAT PROFICIENCY

Dr. Edgar L. Shriver and Sarah Elizabeth Zach

Kinton, Incorporated
Alexandria, Virginia 22311

INTRODUCTION

Training effectiveness is determined by measuring whether or how well the resulting performance meets a predetermined standard. If performance is judged to be adequate, then the training is said to be effective. This assumption relies heavily on the validity of the standard and on the tests by which performance is measured. The purpose of this paper is to investigate how far current standards and tests go towards measuring, and therefore defining, the level of performance we want to achieve through training.

In a recent study of the effectiveness of training for the USA Field Artillery's Fire Support Officer (FSO)it was found that his training did not prepare him for the demands of an intense combat situation, although he was meeting the training standards set for his performance. This meant that, using conventional training effectiveness analysis techniques, current training methods were judged to be adequate. Further study, however, showed that tasks were being performed as a procedural walk-through because the demands of combat were not present. For instance, an FTX or CPX require the FSO to monitor one or two messages an hour; combat demands may be as many as 480 in the same time! The problem was a discrepancy between the training standard and the real world standard of combat proficiency. The training standard was limited almost entirely to the procedural aspects of the FSO's job and did not require the exercise of skills above that level. This is a common deficiency among performance measurement systems today. Many types and levels of skills go into high level performance; an attempt to measure and/or train the complex network of interdependent skills with one simple procedural approach will not achieve results that have any predictive value in the real world.

The approach to performance measurement used in the FSO study was to develop sets of conditions that represent the demands placed on an FSO in an intense combat situation. Then standards were defined that reflect optimum performance under a specific set of conditions. Using these standards of ideal combat proficiency, performance resulting from current training was measured.

The effectiveness of the various methods was evaluated in terms of the percentage of combat proficiency achieved as a result of training. This system of performance measurement differs significantly from most approaches in that the standard it used was the overall product of performance rather than the individual processes.

Constructing the standard required putting together clusters of tasks with assumed combat conditions. This is in distinction to the usual task analyst's proclivity for reducing tasks to constituent subtasks in order to define standards. The reductionist approach makes the ability to perform a subtask the standard for the tasks. This approach is useful in identifying the training content for novice personnel. But, used as a standard for testing, it reduces the test to a measurement of procedural level skill, although the combat skill requirements are actually much higher. The result is tests and test situations that do not rise above the standards for novice personnel. It may not be apparent to military commanders that the tests they use for evaluating performance of presumably highly skilled personnel are merely measuring novice abilities. To draw a fine point, perhaps no one can say the tests do not measure what they purport to measure because no one said they were supposed to measure high skill levels. But on the other hand, it is implicitly assumed that current tests do measure the ability to perform at combat skill levels. So implicitly, if not explicitly, performance tests tend not to measure what they purport to measure.

It is hard to say how we have arrived at the current state of affairs--perhaps we never got above the level of procedures. Task analysis may have just provided us the tools to be very precise about procedural steps. It may be that we can be very objective at the procedural level. Test developers always want to produce objective tests. When we create tests to measure high skill levels we are measuring job products--the accuracy and speed with which they are delivered. This may be inherently dissatisfying in terms of the test developers motivation for objectivity. High skill level tasks, however, are not performed in isolation; they are part of a group of interdependent actions. In complex situations, measuring performance in terms of individual conditions and standards is artificial. In combat, tasks not only must be procedurally correct but also must be performed under time demands, in conjunction with other critical tasks, and with priorities established as to which task will be performed next. The combat performance standard is met if the products are produced, that is, if the desired effect is achieved within the time demands of a specific set of conditions. By measuring performance in terms of the product rather than the process, we are taking for granted the completion of all the procedural aspects of the job; if they were not completed correctly, the product of performance would not meet the standard. But we are adding the demand for high skill level proficiency, not just procedural level "walk-throughs."

It is the thesis of this paper that the inappropriate use of "objective" or process standards to measure complex skills has often resulted in the "proceduralization" of jobs and tests. This in turn has led to a diminished degree of proficiency as a result of training. The cause and effect relationship is obvious--if jobs must be proceduralized to be measured, then training and testing will be conducted at that baby step level. If we want to achieve a higher degree of proficiency as a result of training, we must first make the standards and the tests reflect the skill level desired.

A MULTI-LEVEL PERFORMANCE MEASUREMENT SYSTEM

The first step in constructing any training model is to delineate the job performed and the skills necessary to do it. This is commonly done by means of task analysis. Task analysis was developed originally to deal with the problems of identifying the necessary actions, conditions, and standards of an equipment-structured job. Such jobs, typically maintenance or operational, can be analyzed down to a level of detail that makes the identification of these actions, conditions, and standards on a task-by-task basis relatively easy. However, tasks in many jobs, especially those involving soft skills such as decision-making, cannot be analyzed at this level of detail.

We have not fully recognized that traditional task analytic tools, so useful for producing detailed descriptions of machine-dominant situations, are not always appropriate for analyzing complex combat situations. There are many job tasks or aspects of job tasks that involve so few procedural elements that it is not productive to try to itemize them for training or to use these aspects as a measure of proficiency. Mel Montemerlo used an anecdote to make this point in a paper he presented recently.[1]

Task Objective - To paint a masterpiece

Step 1 - Choose an epic topic
Step 2 - Work out the details in terms of areas to be colored on the canvas
Step 3 - Fill in the areas with color

Almost any task, even the most complex, has some procedural elements. In the example of "to paint a masterpiece" there are strict procedures associated with painting techniques. However, the element that ensures that the result of the task will be a masterpiece is that of genius--a skill in manipulating concepts. The procedural elements of the job--preparing the

[1] Montemerlo, Melvid D., "Politics: The Human Factor in Instructional Innovation." Paper presented to the American Psychological Association, September, 1979.

canvas, setting the palette, etc.--are the ones that allow the genius to be communicated; both types of skills (procedural and conceptual) are necessary to produce the finished painting. When we evaluate a piece of art we evaluate both the "technique" and the "genius." When we train artists, we do not try to train genius--we give them the procedural tools with which to express their genius--if they have any. This is a commonly accepted approach in the arts, and to a large extent in the academic community, but is not widely used in the real world. Here we try to proceduralize even the most complex situations and to measure performance by checking off items on a list.

In order to measure the performance of any person performing a complex job, the different skills for that job must be identified. Then these skills must be classified according to their level. For the purposes of this paper, seven skill levels have been identified. These skill levels do not represent a continuous progression along a continuum but rather a complex jigsaw puzzle with interlocking and overlapping pieces--if you ignore any piece of the picture in any job the description will be incomplete. Some of the distinctions involved in the classification are highly artificial, but they allow us to approach a complex job on many levels instead of just one.

1. Simple Procedural Level

At this level, a verbal or written instruction of what to do is sufficient for job performance. I call this the "stamp licking" level. Most people in our cluture can lick a stamp and put in on an envelope if told to do so. At this level the procedure defines the task and the performance standard.

2. Detailed Procedural Level

This is a class of tasks that can be performed with "how to do it" guidance. The term Job Performance Aid is commonly used to describe such "how to do it" material. The term self-instructional training is also used. By whatever name, the guidance provides graphics and short words describing the actions to be taken. It is assumed that the subject population has the simple procedural level skill to follow each instruction.

3. Complex Procedural

a. Psychomotor

These are tasks that require physical skill linked to some mental activity. Individuals must rehearse the task to achieve the skill needed to produce the product. Tracking a target with a weapon is a common example. Unless the person develops a high level of skill the product is not produced at all. A potential piano player learns certain procedures of where to put his fingers to press out the notes. But then he must practice to develop skill.

b. Representational Verbal/Symbolic Manipulation of Reality

These are tasks that require non-physical skills such as memorization of symbolic codes and manipulation of these languages at a conceptual level. Skills in using maps, symbolic diagrams, etc. are examples of this type of task.

4. Matching (Transfering) Information

This is a class of skills that draws on a person's ability to take information learned in one situation and apply it to another situation. This skill usually comes as a result of training in general concepts or theory. At a high skill level, this is a purely deductive process--drawing on various sources for necessary information.

5. Problem Solving

This skill involves inductive reasoning, e.g., interpreting cues, applying general concepts, and making decisions with incomplete information. Combat leaders involved in tactical operations require this skill. They are usually operating against human opponents who are attempting to manipulate the same situation to their own advantage. The game of chess is a simplified version of a situation requiring these skills.

6. Time Sharing--Any Combination of Skills

At high skill levels, jobs usually require that several tasks be performed at the same time. The job may also require judgments to be made concerning the priority of tasks. Time sharing is a skill that can be taught only in the most general terms--it takes experience to handle many job demands simultaneously. This skill is often found in combination with other higher order skills, seldom in isolation.

7. Invention (Manipulation of Concepts)

This is the highest skill level and is not reducible to procedures. As in "painting a masterpiece," there are some procedural aspects and a great deal of other skills involved. However, at this level we are dealing with the nebulous concept of "genius." This level does not come within the scope of our discussion.

All of the skills described above fall in two general categories--procedural or conceptual. The main distinction between these two categories is the degree to which the outcome of actions can be predetermined. At the lower skill levels, behavior is virtually all "deterministic;" at the highest skill level is almost all "probabilistic." At level 3 we move into the grey area where the proportion of procedural versus conceptual aspects of the job begins to shift. The jobs can still be analyzed down to step-by-step procedures, but the usefulness of this process in terms of performance measurement decreases. It is at this level that practice begins to play an increasing role in proficiency. As the balance shifts from procedural to conceptual, it becomes difficult, if not impossible, to get a true measure of proficiency based on the performance of individual tasks. By recognizing that some skills are procedural and some are conceptual, we should not fall into the trap of trying to measure them all with a simple test of procedures or even to analyze them with the same techniques.

Having identified the need for a multi-level performance measurement system, we must now discuss how to achieve it. Like so many good ideas, that is easier said than done. Procedural tasks can be measured with procedural checklists; slightly more complicated tasks can still be proceduralized and measured at that level. It is at the point at which the job cannot be usefully proceduralized (or at least it would not be productive to do so) that a new form of performance measurement is needed. I believe this measurement system will have two major elements: (1) standards that measure the product rather than the process, and (2) a testing situation that will exercise all the aspects of the job.

We spoke earlier of "product" measure that presume the correct performance of the procedural aspects of the job and described how combat proficiency standards were developed in the FSO study. Using these kinds of standards involves a complete "about face" from the traditional checklist approach; that is, instead of reducing tasks to their elements we must cluster the tasks that contribute to the production of a specific output. Making the initial change in focus from process to product is the hardest step; after that it will be relatively easy to develop training and testing situations that achieve the desired results. The additional element that is necessary to affect this change is the development of new analysis techniques for complex, non-procedural jobs.

I suggest that the "new" analytic techniques will not be deterministic. Traditional task analysis is deterministic. Their use assumes all stimuli responses and conditions can be stated in such a way that when properly performed, standards will be met. I suggest that our next evolution of analytic techniques will be probabilistic. I believe they will be used to create training situations in which participants can practice actions with uncertainty regarding input cues and uncertainty regarding the results of their actions. These types of probabilistic situations have a great potential impact on the way people learn to perform mid to high skill level jobs. Training will no longer be restricted to a procedural walk-through but will offer the possibility of fully simulating an actual job environment.

The type of analysis that is necessary to produce these kinds of sophisticated training situations is one that identifies the contingencies of the job as well as the procedural actions, conditions, and standards. A contingency analysis would be a way of describing what is behind the scene presented to the job performer. Using traditional task analysis we have done very well at identifying and proceduralizing the "inner workings" of jobs such as electronic troubleshooting. The new techniques should be able to describe the contingencies of a situation created by humans rather than by a machine.

It is important to remember as we attempt to analyze complex conceptual jobs that humans do not act in the same logical, predetermined way that machines do. In chess, or business, or combat, people make decisions based on their own perceptions of the situation. To train them to do this more effectively, we should not try to proceduralize the job (which would be impossible) but rather to give them an understanding of the "inner workings" of the process they face. This kind of analysis goes well with a "product" approach to performance measurement since it puts its emphasis on the results of actions rather than on the individual processes used to achieve them.

A workable measurement system that demands high level proficiency is the sine quo non of increased training effectiveness. It has often been said that if you control the test, you control the training. The model described above provides an outline that can be used to develop both the tests and the training that will ensure increased performance proficiency.

DIFFICULTY
WHY IS IT SO HARD TO TALK ABOUT IT?

By
John P. Smith

When we say that some subject is difficult we mean it is hard to learn. When we ask why it is hard to learn we usually hear only three answers.

SLIDE 1

(1) The subject is complex in that it requires many facts, ideas, or concepts to be understood or kept in mind at one time. (2) Abstract subjects are believed to more difficult to learn than concrete subjects. (3) If the emphasis shifts from the subject matter to the learner, then certain individual differences are said to be related to ability to learn.

These traditional explanations are not wrong, but they're not very useful. They suggest no way to deal with difficulty save by selecting people whose attributes compensate for complexity and abstractness. We don't have perfect selection, so we deselect, by academic attrition, those who do not compensate. We also accept as a further cost the additional years of on-the-job experience needed to develop real competence. It seems to me that neither Learning Theory, Educational Psychology, nor Instructional Technology give us much help in talking about difficulty.

So, I tried another window through which to look at the problem. My thesis is that there is a fourth factor in difficulty: In some courses failures occur because of inadequate training materials. There are clues to the preparation of good training materials if we search for them. I can share with you some of the results of my search.

I will not cover remedial and basic skills work, as this very important research has a different emphasis. I omit discussion of the many commercially successful multi-media developments because, to the extent they are motivated by learning problems they prefer to deal with them by adding resources at some cost, not by reducing difficulty. I prefer the alternative, reducing difficulty without adding cost. My approach is to identify good practices in the Instructional Technology literature, and then look into some of our courses to see how they compare.

Inappropriate Use Of Discovery Techniques

Bunderson and Dunham (4) carried on a lengthy research program on cognitive abilities and learning. One aspect of their work concerned the interaction of reasoning and memory abilities with "discovery" and "rule-example" methods in the learning of an imaginary science called "Zenograde." The discovery groups required more time and more examples, but didn't perform better and were inferior on some indices of retention and transfer. Students lower on reasoning and memory abilities had considerably more difficulty with the discovery treatment than with the expository presentation.

Inappropriate use of discovery techniques may occur if training materials attempt to teach complex subjects by example and not by rule or structured exposition. I reviewed some technical course materials to see if I could find any instances where this occurred. As it turned out, I didn't have very far to look. The topic, "Variational Analysis," involves the simultaneous alteration of quantities with both direct and inverse relationships--for example, what happens to branch and total current when resistance is increased or decreased in some part of a series-parallel circuit. The teaching materials only gave a few examples, with no specific rules or procedures. These exhibits essentially comprise the entire guidance for this important, integrative, intellectual task.

SLIDES 2-8

The greatest number of failures in this course, 22%, occur on this topic, and most of the unmotivated, unable, and math-deficient students are eliminated before this point. No doubt other factors contribute to this failure rate, but the Instructor and Counselor comments in the failed students' jackets implicate variational analysis in a great many cases. Would the students do better if some sort of procedural guidance were given? We don't know; it hasn't been tried. It does seem plausible that students low on reasoning and memory would be the ones to have the most trouble. At any rate, we have the suggestion that inappropriate use of discovery presentation techniques is a cause of needless difficulty.

Topic Organization

Bonnie Meyer (13) noted that, of two passages from The Scientific American, one was recalled nearly twice as well as the other by subjects in a learning experiment. She also observed what appeared to be different types of organization in these passages. Subsequently, a passage on "loss of body water" was written in four different forms of organization, containing identical "idea units." The organization form called "cause-effect" (covariance) and the form called "description of qualities," (attribution) were different in the number of idea units recalled.

SLIDE 9

This data suggests that we need serious study of the effects of topic organization on learning and retention. Merely taking the notion that students pick up on cause-effect statements, I looked at some materials where such statements would have been possible; that is, in simple electrical system troubleshooting. Unfortunately, I wasn't able to reach any conclusions that would relate to Meyer's work, as the material had other characteristics which appeared even more powerful; descriptive material was incomplete, dispersed in different modules, and phenomena and procedures were confounded with each other and with other subject matter.

Density of Subject Matter

The Bell Telephone Company has a huge training job, perhaps second only to the military in the number of people trained every year. Frase and Schwartz (6) at the Bell Laboratories, analyzed some Bell training material. One interesting finding was, "12% of the sentences (in a particular text) caused problems because they contained several meaningful components, such as actions to be taken, conditions for those actions, parts of equipment, causes and effects, and so on." "Segmenting" of material produced a 12% saving of learning time. Segmenting plus "indenting" saved 18% of the learning time.

SLIDE 10

For the question of difficulty, a possible clue is the reference to the density of meaningful components, as high density may overload processing capability. Here is an example from one of our Navy courses.

SLIDE 11

Segmenting and indenting appear to be of some benefit in mitigating density, but this passage may be beyond this remedy. The last line makes the entire passage questionable; technicians do not work with numbers like these.

<u>Set</u>

The idea that the student's mental set has a marked effect on learning is as old as Psychology. We attempt to provide a set by some sort of pre-instructional orientation such as an overview or an objective. However, we perhaps should give more thought to this. John Bransford and students (3) studied orienting instructions in an "imaging" task, using a photograph of a room.

SLIDE 12

Note the differences in the number of objects recalled as a result of the different orienting instructions.

Pichert and Anderson (16) gave two groups different instructions before reading the same description of a house. One group of subjects was told they were to burgle the house, while the other subjects were instructed that they were potential home buyers. On a later test, there was very little overlap in the topics recalled by the two groups.

With other students, Anderson (2) had passages that could be read as a description of a prison break or as a wrestling match, and as a card game or as the rehearsal of a woodwind ensemble. Subjects were Physical Education or Music Majors, and you can guess how these groups interpreted the passages. Anderson states, "Our main thesis is that the meaning of a communication depends in a fundamental way on a person's knowledge of the world and his/her analysis of the context as well as the characteristics of the message." (p. 368) I add that the person's "analysis of context" does not necessarily occur as a deliberate, aware process. Rather, it may be inadvertent and implicit.

We can also look to the scrambled word game for illumination of the effects of set. In this game, the letters of some ordinary word are scrambled, and the task is to rearrange them and rebuild the meaningful word. Here are a couple of examples.

SLIDE 13

You go ahead and try to identify the words. Do not say them aloud, but show a hand when you get them both. I will keep time and see how long it takes. (Pause) It may help you to know that these are common food items. (Pause) The words are "cheese" and "orange." Now, try the next two words.

SLIDE 14

(The audience will attempt to make food words; this will not succeed.) Are you trying make foods out of these jumbles? Oh, I forgot to tell you, these are not foods, they are common tools. Hand tools. (Pause.) I reiterate Anderson's point: The learner provides a context, sometimes an inappropriate one, and misdirection may easily happen.

Advance Organizers

David Ausubel's concept of the "advance organizer" as a means of enhancing learning has been around for 20 years without getting much research support. Recently, Richard Mayer, (12) at the University of Califronia at Santa Barbara, identified five characteristics of effective advance organizers and four conditions under which they may be effective. One of the requirements is that the organizer must influence the student's encoding process, either by connecting the training input to the organizer or by relating the input to some structured knowledge the student already possesses. That is, the advance organizer must either provide structure or activate existing structure. If it merely points out what the topic is about, it cannot be effective. Thus, the statement, "This topic concerns the relationship between current, voltage, and resistance in a series circuit," could not help the student who did not already have these concepts as a result of previous learning. For the total novice, something like the old hydraulic analogy used many years ago might be useful.

The What, Why, and How of Technical Training

In our most effective and most interesting courses, what we teach is major course objectives, why we teach them is because they are mission relevant, and how we teach them is in a manner to produce a coherent performance. In other courses there is a problem of selection of content, which also influences organization and teaching method. These courses seem to be trying to teach a vast store of details such as facts, nomenclature, technical data, definitions, concepts, and rules. Why we are doing this is not clear. It seems to be assumed that the student will need all of this information in his future work, and that he will remember it and retrieve it as needed. Richard Anderson (1), for one, does not accept this approach.

> "On the naive view that the text contains the meaning, the words themselves are valuable. The student's capacity to recognize or reproduce them accurately is evidence that he or she possesses the knowledge the text conveys." (p. 423)

The nature of the material limits and directs the choice of instructional method, so how we teach the material is by a detail by detail, topic by topic progression. Here is a sample of a frame from such a linear sequence on magnetism.

SLIDE 15

The frame concerns one of the characteristics of magnetic lines of force. Similar frames on the other characteristics are given.

The information is tested:

SLIDE 16

Many courses use this type of recognition test item to monitor student progress. Perhaps we use more test items in our CMI courses than in the traditional classes, because we believe tests are essential to self-pacing. But our practice sometimes is not very convincing. We want to know about the student's mastery of important objectives, yet we give some test items some of which concern only insignificant details. Several questions can be asked of such material: Does effective learning management really need the detailed testing we now use? And, where is the larger meaning of passages like this on magnetism, what is the student enabled to do when he completes it? Might we inadvertently shape the student so that he believes he has learned the material when we have only assured a transitory recognition?

Perhaps magnetism isn't a very important subject with which to demonstrate these points. The same questions can be asked about material that is more clearly relevant to technical work, such as the use of test equipment.

SLIDE 17

An extensive text, of 67 frames, provides pictures and diagrams of the multimeter and directions on how to set, adjust, and read the meter. These frames fairly represent this material.

SLIDES 18-20

This looks like useful information, and most of us would probably judge this lesson to be good training material. Yet, Lab Instructors tell us that many students who pass the written test on this material do not know how to use or read the meter a short time later. We do not know with certainty why this difficulty occurs. Perhaps the Lab Exercise instructions are inadequate. Perhaps criterion performance under the linear approach is not sufficient to ensure retention for even a day or so. Modern Cognitive Psychologists believe that retention and retrieval are largely determined by the organization of material when it is learned. Perhaps the problem is that the serial progression lacks the organized, meaningful structure in which the multimeter testing procedures, discriminations, and decisions find an anchorage that is tapped by the later applicational task.

Donald Norman (14, 15) argues that the linear progression is very vulnerable to disruption because there is only a single chain of associations. He argues that anything which must be remembered should be incorporated into a "web of associations" which has many pathways for retrieval. Here is an illustration of what some electronics material might look like in Norman's suggested organization.

SLIDE 21

Each line of the diagram is as important as the nodes, and represents a training objective just as the nodes do. The technical data and information on the use of the meter does not appear on this "web" and would not be taught until after this structure was understood. It also seems to me that the lines of this diagram could be used as a heuristic to remind course writers of some of the points they should include in discussion of topics.

Behavioral objectives also need some discussion in the context of what, why, and how we teach. A review by Duchastel (5) indicates that behavioral objectives often do not facilitate learning.[1] When they are helpful it is because of their directive effect; the student is cued to attend to certain information and disregard the rest. There is a sort of harvesting analogy here; as if the student was supposed to sift through the chaff in search of the wheat. But why confront the student with a pile of chaff in the first place? Why shouldn't the behavioral objective provide guidance for the course writer so he produces wheat instead of chaff?

Perhaps what sometimes occurs in the development of some of our military courses is that the course writer sets down what he knows or thinks is important about a topic, electron theory for example, and then writes the behavioral objective.

SLIDE 22

The student will label the electron, the neutron, and the proton, given an unlabeled diagram, without aid, and with 100% accuracy. This is trivial, which makes the standard inappropriate, and the job relevance is doubtful at best.

Knowledge analysis to identify significant, not trivial, subject matter is rapidly gaining recognition. Robert Glaser (7), writing of the development of competence, says, "There seem to be two main elements (to the analysis of competent performance). One is identification of the information structures that are required for performance, and the other is a description of the processes and cognitive strategies--heuristics and algorithms--that need to be applied to this information <u>and which themselves are part of the information data base</u>." (Emphasis added.) And Ernst Rothkopf (17) of the Bell Telephone Laboratories, says, "A decade of experimentation in the public schools and in military and industrial training indicate strongly that the most important results have not been produced by changes in instructional materials, nor by use of computers or audio/visual devices, but <u>rather by changes in instructional content</u>." (Emphasis added.)

Harmon (8) suggests that we sometimes confuse motivational and instructional tasks, and fail to differentiate material which needs operant or cognitive instructional techniques. He has developed an ISD algorithm that could well be the outline for the next phase in the evolution of our current ISD model.

Summary

Today I have discussed the idea that we may create learning difficulties for some students by the way in which we write training materials. Use of discovery methods instead of algorithms, certain types of paragraph structure, density of substantive elements, and students' mental set may all create

[1] Other reviewers also found no benefit for behavioral objectives (see, e.g., Lawson (10) and Hartley and Davies (9)). One of the more complete reviews, as well as the more critical, is that MacDonald-Ross (11) whose 1973 review lists 16 objections to their use.

difficulties. Some of our training includes trivial objectives, not because anyone wants them but because they follow from the conception that the goal of training is to fill the heads of students with a vast store of highly detailed, and sometimes unrelated, information. This creates a requirement for brute force memorization, because there is no meaningful structure to which the detail can be assimilated. Retention and retrieval for later application are dubious.

We seem to have gone to a lot of trouble with our behavioral objectives for questionable return; perhaps they would be more functional if they were considered to be directive guidance for writers instead of for students. The remedies for these difficulties that we have created for students are to increase our ability to use knowledge analysis to identify important and worthwhile subject matter, as Glaser and Rothkopf suggest, and to emphasize relationships in subject matter as Norman indicates. These can be accomplished if we have the goal of creating some set of coherent accomplishments, and not merely passing out a lot of information. We may also be learning how and when to use advance organizers. And Harmon's very interesting algorithm may clarify many decisions for course writers. Looking at the rapid recent evolution of Instructional Technology, we might hope that we will have less difficulty to talk about, and less difficulty talking about it, in the future.

REFERENCES

1. Anderson, Richard C., "The Notion of Schemata and The Educational Interprise," In Anderson, Richard C., R. J. Spiro, and William E. Montague (Eds.) Conference Proceedings: Schooling and the Acquisition of Knowledge, San Diego, Naval Personnel Research and Development Center, December 1977 (TR 78-6).

2. Anderson, Richard C., R. E. Reynolds, D. L. Shallert, and E. T. Goetz., "Frameworks For Comprehending Discourse," American Educational Research Journal, 14, 4 (Fall 1977), 367-381.

3. Bransford, John D., Kathleen E. Nitsch, and Jeffrey J. Franks, "Schooling and the Facilitation of Knowing." (In Anderson, et al)

4. Bunderson, C. Victor, and J. L. Dunham, Research Program on Cognitive Abilities and Learning, Final Report, Austin, The University of Texas, CAI Lab, 1970.

5. Duchastel, Phillippe, and Paul F. Merrill, The Effects of Behavioral Objectives on Learning: A Review of Empirical Studies," Review of Educational Research, 43 (1973), 1, 53-69.

6. Frase, Lawrence T., and Barry J. Schwartz, Typographical Cues That Facilitate Comprehension, Journal of Educational Psychology, 1979, 71, 2, 197-206.

7. Glaser, Robert, "Components of a Psychology of Instruction: Toward a Science of Design," Review of Educational Research, 46 (1976) No. 1, 1-24.

8. Harmon, Paul, "Beyond Behavioral Performance Analysis: Toward a New Paradigm For Educational Technology, Educational Technology, 19, 2, 5-26 (February 1979)

9. Hartley, James, and Ivor K. Davies, "Preinstructional Strategies: The Role of Pretests, Behavioral Objectives, Overviews, and Advance Organizers," Review of Educational Research, 46 (1976) No. 2., 239-265.

10. Lawson, Tom E., "Effects of Instructional Objectives on Learning and Retention," Instructional Science, 3 (1974), 1-22.

11. MacDonald-Ross, Michael, "Behavioural Objectives: A Critical Review," Instructional Science, Vol. 2 (1973) 2-52.

12. Mayer, Richard E., "Can Davance Organizers Influence Meaningful Learning?" Review of Educational Research, 49, 2, 371-383 (Summer 1979)

13. Meyer, Bonnie J. F., "The Structure of Prose: Effects on Learning and Memory and Implications for Educational Practice," in Anderson et al.

14. Norman, Donald A., "Memory, Knowledge and the Answering of Questions," in Salso, R. L. (Ed) Contemporary Issues in Cognitive Psychology, Washington, D.C., V. H. Winston and Sons, 1973.

15. Norman, Donald A., Donald R. Gintner, and Albert L. Stevens, "Comments on Learning Schemata and Memory Representation," in Klahr, David (Ed) Cognition and Instruction, Hillsdale, N.J. Lawrence Erlbaum Associates, Publishers, 1976.

16 . Pichert, James W., and R. C. Anderson, "Taking Different Perspectives on a Story," Journal of Educational Psychology, 1977, 69, 4, 309-315.

17. Rothkopf, Ernst, "The Concept of Mathemagenic Activities," Review of Educational Research, Vol. 40, (1976) No. 3, 325-336.

TASK SELECTION FOR JOB PROFICIENCY AND TRAINING

Michael J. Cassidy, Hendrick W. Ruck, and Stephen V. Offutt

Air Force Human Resources Laboratory (AFHRL)
Brooks AFB Texas 78235

INTRODUCTION

The on-the-job training (OJT) system used in the Air Force relies on two different sources of instruction. Career-wide knowledge and background are provided via standardized career development courses (CDC). Completion of these CDCs is mandatory for being upgraded from the apprentice to the journeyman skill level. In addition, hands-on training must be conducted and certified by supervisors. Specialty Training Standards (STS) are used as guidelines for this hands-on-training. Many supervisors have reported problems in identifying the tasks on the STS that should be trained (Stephenson & Burkett, 1975). Supervisors generally agree that training all STS tasks is impractical, due to equipment and time availability, and training only one task is usually insufficient. However, no system for identifying the "important" job tasks for hands-on OJT is operationally available. This paper reports on the tentative framework of such a system, and the research questions and design being applied to that system.

One of the key problems encountered when using STSs as guides for OJT is that the STS is specialty oriented, not job oriented. Since the hands-on component of OJT is job oriented, a system for annotating the STS to make it job-relevant is necessary. In the present system, supervisors are free to annotate the STS and add to it for job specific training tasks. However, supervisors have indicated that they prefer some guidance in this annotation process.

BACKGROUND

The Standardized Position Oriented Training (SPOT) concept has been developed in response to the above mentioned problems. The SPOT concept suggests that standardized training task lists for each job in a specialty should be developed for use in guiding hands-on OJT. The primary features of SPOT are (a) tasks are listed by job, (b) training task lists are standardized, and (c) task lists are used instead of the more general STS for hands-on OJT.

SPOT is not a system that will be different in kind: it is a major modification of the existing hands-on OJT system. It will provide standardized guidance to the field supervisor for the decisions he or she is now required to make. The broad, generalized statements of the STS covering the entire specialty will be replaced by a series of task lists each identified with a specific job. The tasks will be those describing the important core of the job. Before a supervisor would certify an airman as competent to work at the job at any skill level, he would first check the airman's proficiency on each of these core tasks.

Flexibility is inherent in the SPOT system because it provides both guidance and direction. Local conditions, such as aircraft type, MAJCOM maintenance requirements, equipment, and manning, may result in particular jobs varying in task content from the SPOT specified job. Therefore, the supervisor will be able to modify the SPOT list to match a particular job by deleting and adding tasks. This flexibility is the key to the utility of standardized guidance. Once the supervisor has a modified (if necessary) SPOT list which describes the core of a particular job, he would proceed to check an airman's proficiency on all tasks on the list. This process will define the extent of hands-on OJT required to certify the airman as competent at that job.

This system of guidance differs from the present system (Stephenson and Burkett, 1975) in two essential respects. The STS is replaced by a number of task specific SPOT lists which are necessarily performance related listings. To develop SPOT lists, the occupational survey job inventory is screened to produce a list of core tasks for each job in the specialty. The time-consuming, decision-making process of reviewing the STS, selecting training tasks, and defining and adding tasks not covered by the STS will be shortened so that the supervisor will increase the proportion of OJT-time spent on the mission-enhancement, hands-on OJT and decrease the proportion spent on administrative paperwork. A significant improvement in the SPOT system over the present system will be the inclusion of mandatory skills and knowledge (as defined in the Airman Classification Manual, but not covered in the STS) for each skill level of a specialty. Provision will also be made for contingency tasks, which are MAJCOM specific or combat related, to be included. This will enhance the role of OJT in assuring operational readiness. Tasks which are important for safety, but do not meet the occupational survey data selection criteria, can also be included in review by Air Staff or MAJCOM functional personnel.

In developing the SPOT system, some of the research questions to be answered are: (a) how should jobs be identified, (b) how should tasks be selected, (c) would additional information, such as task analyses, be significantly more useful than only task lists, (d) what percentage of tasks should be trained, and (e) what is the impact on unit effectiveness and training load of the SPOT system. Of these questions, this paper will focus on the task selection process. As you will see from the discussion, the question of what percentage of tasks should be trained is answered by the selection rationale: proficiency will be required on all listed tasks for a job; OJT will be required on those tasks on which an airman is not proficient on an initial check. The issue of supplementing SPOT lists with task analyses of the listed tasks is being addressed in the research project, but is outside the purview of this paper.

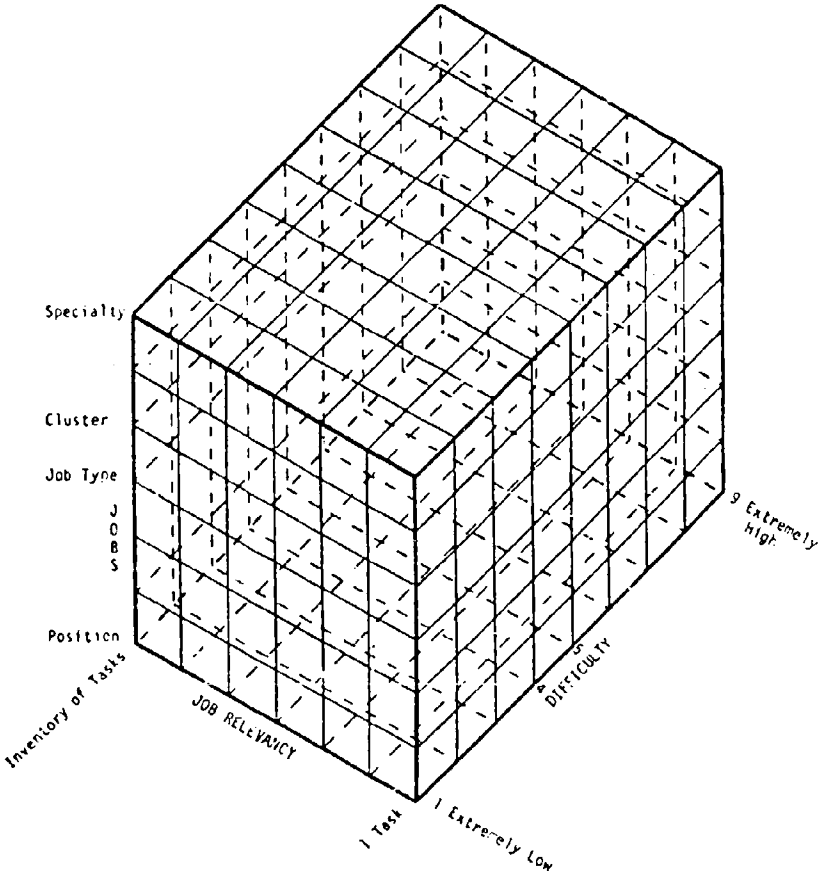
Initially, the SPOT lists will be developed from occupational survey data that have been collected as a part of the Air Force occupational survey program. The major initial research question to be answered, then, is, "How should tasks be selected for inclusion in standardized hands-on OJT?"

TASK SELECTION PROCESS

We have imposed as a primary constraint upon selecting tasks for inclusion in SPOT lists the use of data from the Air Force occupational survey program. Why should this constraint be imposed? As already noted, the SPOT system is a response to the 'preference' of supervisors for guidance in the task-selection-for-OJT process. There is an OJT system that is operational, so the cost of any change would be subjected to the closest scrutiny. In this environment, the automated manipulation of data from an operational program presents the highest cost-benefit product available. Hence, by accepting this primary constraint, the greatest potential for research product implementation is established. What, then, does the occupational survey program offer for the task selection process?

There are four categories of information available which could contribute to the selection of tasks for inclusion in standardized hands-on OJT: (a) job identification; (b) relative time spent; (c) task difficulty; and (d) task training emphasis. The last of these, task training emphasis,

Figure 1. DOMAIN OF SPOT



was discarded because of its inherent conflict with the objective of SPOT. Training emphasis, as collected in the occupational survey program, is defined in relation to the emphasis needed to train first term airmen. SPOT is conceived as applying to the whole specialty, regardless of skill level. Therefore, to use training emphasis data would bias the SPOT lists towards first termers; and, potentially, all tasks not performed by first termers would not be included. So by eliminating training emphasis, a viable, three-dimensional data base will be utilized for our task selection process (see Figure 1).

Job identification describes the dimension jobs, a discrete scale ranging from a position, through a job type, a cluster, to a specialty. The position job is the group of tasks performed by an individual; the specialty job is the family of tasks performed by members of an Air Force Specialty (AFS), in other words, the aggregate of all position jobs within an AFS. Along this scale, we can find the job types and clusters reported in Occupational Survey Reports (OSRs). These levels of jobs are groupings of positions with significant similarity, both of the tasks performed and the relative percentage of time spent on those tasks. Other groups of positions could be identified on this job scale. However, to require such identification in the operational implementation of SPOT would complicate the data manipulation and could confuse the information presented to management in the Occupational Survey Reports.

In opting to use job types and clusters identified in OSRs, the implementation of SPOT will be simplified. The risk inherent in this approach is that the quality of the task inventory development and the occupational analysis become crucial to the utility of the SPOT lists for the field supervisors. This risk is acceptable for now because use of occupational survey information in other Air Force programs indicates that the information is very usable. Further, the experimental design for testing the utility of the variants of SPOT listings permits an accurate assessment of the impact of specifying jobs at these levels. Depending on research results, additional methods of defining jobs may have to be developed.

Another advantage of using job types and clusters published in Air Force Occupational Survey Reports is that they facilitate classification of Air Force Specialties as homogeneous or heterogeneous. We anticipate that compiling SPOT lists for heterogeneous specialties (for example, those with more than 15 job types) will be more difficult because of the greater specialization of positions. The supervisor's task of matching a particular list to a given position could be considerably more difficult due to the number of SPOT lists, and perhaps, a greater potential for lack of fit. The inclusion of both homogeneous and heterogeneous specialties in the experimental design will allow for the assessment of the general applicability of SPOT.

The second dimension, job relevancy, is based on relative time spent data for this first cut of task selection. As a result of this research, additional factors may be added. The job relevancy of a job ranges from the one most "critical" task in a position through to all tasks performed in a specialty. The description of the job centers on one task and is expanded by adding those tasks which occupy most of the time of all members performing the job, that is, tasks which are commonly performed by members in the job, taking relatively more time than other tasks, and/or are performed more frequently than other tasks. So, through job relevancy, we are looking for those tasks on which a supervisor would require an assurance of proficiency before accepting an airman as competent in a position. Operationally, the best vector for describing job relevancy appears to be the "cumulative sum of average percent time spent by all members" in a job type because it embraces the commonality, the time taken, and the frequency of performance of tasks. (Other factors which could contribute to job relevancy as a discriminant of tasks are: consequences of inadequate performance, task delay tolerance, hazard potential, training emphasis, and interactions among these vectors. These vectors will be investigated in the policy-capturing phase of refining the first cut task selection process. This structured selection process is the main thrust of the task selection research to insure that the procedural model developed is not unnecessarily complex).

In the research on developing SPOT task selection methods, all tasks performed by incumbents in a specific job type are ordered on the average percent time spent by all members. (The raw relative time-spent ratings are converted to percentages to permit comparisons between incumbents across tasks. The composite job description is then produced by averaging the percentage time spent on each task by all members of the group (Archer, 1966). Tasks are listed in descending order. Then, by cumulative addition of the ordered average percent time spent by all members, the most time-consuming (or central to the job) tasks for the job type can be identified. By specifying various cumulative sum cut-offs, the list of job relevent tasks in the job can be narrowed or broadened. Two screens will be tested for job relevancy: one accounting for 50% of the time spent by incumbents in a job type or cluster; the second for 75%. In this way, the tasks commonly performed in a job type or cluster will be considered for inclusion while those which do not take much of the incumbent's time and/or are performed by only some of the incumbents will be screened out.

The final dimension to be used to screen tasks is task difficulty. Task difficulty is defined as the learning difficulty of a task; the greater the time needed to learn a task the more difficult the task is. Again, two screens will be tested: the first allowing into the SPOT lists those tasks with a standardized mean difficulty of 5 or greater; the second, going one standard deviation below the mean to those tasks with a difficulty of 4 or greater. The difficulty screen is included to capture those tasks which would probably require OJT or, at least, on which the supervisor would want to check for proficiency.

In applying the screens for job relevancy and difficulty, four variants of the SPOT list have been produced for each job identified. (These lists are not definitive; they simply expedite the OJT process by giving, for each job, preliminary task lists which will guide the knowledgeable supervisor in making logical and sensible decisions about the proficiency requirements of jobs). The sequence of applying the screens is important. A review of some indicated that the most difficult tasks are performed by relatively few incumbents and consequently made only small contributions on the average percent time spent vector. From this we can conclude that to apply the difficulty screen first would bias the sample of tasks to be considered for job relevancy. This is undesirable as we have chosen to go into the specialty at the job type/cluster level and hence seek to cover as many positions as possible. To allow for the more difficult tasks in a position which do not get through the job relevancy screen, the field supervisor is given freedom to add tasks to a list as his knowledge of the job and task and his experience indicate.

What type of product results from these selection processes? In the Cable and Antenna System Installation/Maintenance career field, a group of 91 airmen from a survey sample of 552 were identified as Antenna Installers and Maintainers. Two of the SPOT lists for this job type (produced from an inventory of 409 tasks) illustrate the selection product (see Tables 1 and 2). The first SPOT list results from the most stringent screening - 50% on job relevancy and 5 on difficulty. Eleven tasks are accepted, accounting for 11.70 of the cumulative average percent time spent by all members in the group. Coupled with the range of 25.3 to 95.6% of group members performing these tasks, these data indicate that a narrow, common core of the job has been selected. The mean difficulty of these tasks is 5.66 with a maximum difficulty of 7.48. This above average to high difficulty supports the need to check proficiency (the result of training) on these tasks. By way of comparison, the least stringent screening - 75% on job relevancy and 4 on difficulty - accepted 84 tasks that accounted for 63.00 of the cumulative average percent time spent. Tasks on this second list are performed by 13.2 to 95.6% of group members. Hence, this list is a much broader segment of the job. While the mean difficulty is now 4.92, there is still an apparent need to check proficiency on all these tasks. These figures indicate that the selection process produces SPOT lists with a good spread over job relevancy at the same time capturing tasks of medium to high difficulty. The percentage of group members performing the tasks indicates that the SPOT lists are potentially applicable to most of the positions in the group. The next phase of the research is to evaluate the utility of SPOT lists for the field supervisor.

Table 1. SPOT TASKS SELECTION

Time Spent 50%, Task Difficulty 5, (Most Stringent Screening) Antenna Installers and Maintainers

D	TSK	Titles	SEQ NUM #	LRN DIF -F-	% MEM (M)	% TIM (T)	CUM % <--
M	1	Compute Voltage Standing Wave Ratios	1	7.48	25.3	.73	.73
K	16	Splice Coaxial Cables	2	6.28	31.9	.56	1.29
P	4	Install or Maintain Pressurization Systems on Waveguides	3	5.97	31.9	.70	1.99
K	7	Install Coaxial Connectors	4	5.73	63.7	1.34	3.33
G	1	Climb Cable Support Structures or Poles	5	5.57	57.1	1.47	4.80
K	14	Remove or Replace Coaxial Connectors	6	5.51	57.1	1.15	5.94
L	23	Inspect or Adjust Timers or Photoelectric Cells	7	5.42	30.8	.71	6.65
D	9	Demonstrate Operation of Equipment	8	5.21	27.5	.58	7.24
D	6	Conduct On-The-Job Training (OJT)	9	5.09	30.8	.67	7.91
L	1	Check Plumb of Antenna Supports	10	5.04	51.6	1.11	9.02
L	2	Climb Antenna Supports	11	5.00	95.6	2.67	11.70

$\bar{X} = 5.66$

Table 2. SPOT TASKS SELECTION

Time Spent 75%, Task Difficulty ≥ 4 (Broadest Screening), Antenna Installers and Maintainers

	D	TSK	Titles	SEQ NUM #	LRN ·DIF ·F·	% MEM (M)	% TIM (T)	CUM % <--
+	M	1	Compute Voltage Standing Wave Ratios	1	7.48	25.3	.73	.73
+	K	16	Splice Coaxial Cables	2	6.28	31.9	.56	1.29
	B	24	Prepare Airman Performance Reports (Apr)	3	6.17	19.8	.31	1.59
	B	30	Research Procedures to Resolve Technical Problems	4	6.11	23.1	.43	2.03
	C	10	Perform Final Inspections on Maintenance of Antennas or Cable Systems	5	6.05	20.9	.39	2.42
+	P	4	Install or Maintain Pressurization Systems on Waveguides	6	5.97	31.9	.70	3.12
+	K	7	Install Coaxial Connectors	7	5.73	63.7	1.34	4.46
	M	4	Fabricate Wire Antennas	8	5.71	35.2	.41	4.87
	A	7	Develop Plans for Performing Maintenance	9	5.68	23.1	.32	5.19
	P	1	Assemble or Disassemble Waveguides	10	5.57	18.7	.32	5.51
+	G	1	Climb Cable Support Structures or Poles	11	5.57	57.1	1.47	6.97
+	K	14	Remove or Replace Coaxial Connectors	12	5.51	57.1	1.15	8.12
	M	16	Sag and Tension Wire Antennas	13	5.49	30.8	.39	8.51
	L	16	Fabricate or Install Guys and Anchors	14	5.46	35.2	.41	8.91
+	L	23	Inspect or Adjust Timers or Photoelectric Cells	15	5.42	30.8	.71	9.63
	M	10	Install Wire Antennas	16	5.36	27.5	.34	9.97
	C	12	Perform In-Progress Inspections During Maintenance Activities	17	5.36	16.5	.41	10.38
	L	31	Install Safety Climbing Devices	18	5.35	28.6	.43	10.81
	P	3	Inspect Rigid Waveguides	19	5.30	29.7	.51	11.32
	G	43	Remove or Replace Aerial Cables	20	5.21	17.6	.30	11.63
+	D	9	Demonstrate Operation of Equipment	21	5.21	27.5	.58	12.21
	L	45	Remove or Replace Guys and Anchors	22	5.20	27.5	.35	12.56
	P	2	Inspect Flexible Waveguides	23	5.18	26.4	.46	13.04
	E	1	Annotate Job Completion Documents	24	5.11	13.2	.31	13.35
+	D	6	Conduct On-The-Job Training (OJT)	25	5.09	30.8	.67	14.02
	C	7	Insure Compliance with Technical Order (To) Specifications	26	5.06	22.0	.49	14.51
	B	13	Implement Safety Programs or Procedures	27	5.05	14.3	.36	14.87
+	L	1	Check Plumb of Antenna Supports	28	5.04	51.6	1.11	15.98
	L	25	Install Antenna Support Crossarms	29	5.03	26.4	.33	16.31
+	L	2	Climb Antenna Supports	30	5.00	95.6	2.67	18.99
	B	19	Maintain Maintenance Data Records	31	5.00	19.8	.43	19.42
			⋮					
	H	32	Test Manholes for Toxic Gases	39	4.86	17.6	.49	27.91
	K	12	Remove or Replace Aerial Coaxial Cables	40	4.86	26.4	.35	28.26
	H	31	Test Manholes for Combustible Gases	41	4.85	17.6	.49	28.74
			⋮					
	L	53	Test Guy Tension	64	4.46	71.4	1.50	48.93
	L	29	Install Lightning Protection on Antenna Support Poles	65	4.45	29.7	.37	49.31
	L	24	Inspect Steel Pedestals	66	4.44	31.9	.86	50.17
			⋮					
	L	36	Paint Antenna Supports	81	4.06	33.0	.52	60.32
	L	22	Inspect Obstruction Markings	82	4.04	35.2	.93	61.25
	C	6	Inspect Vehicles for Condition or Serviceability	83	4.02	56.0	1.41	62.66
	B	36	Schedule Work Assignments	84	4.02	14.3	.34	63.00
					$\bar{X}$ = 4.92			

+ Denotes tasks on the 50% - 5 SPOT List

PLANS FOR EVALUATION OF SPOT SYSTEM

To test the general applicability of the SPOT selection process, the experimental design calls for field testing in a variety of specialties and major commands. Different screening procedures may well be required for various types of specialties. For this reason, SPOT lists have been compiled for eight specialties, two in each of four categories (see Table 3).

Table 3. SPECIALTY CLASSIFICATION FOR SPOT

Homogeneity Technicality	Homogeneous	Heterogeneous
Hard	423X0 - Aircraft Electrical System Career Ladder 361X0 - Cable and Antenna System Installation/ Maintenance Career Ladder	461X0 - Munitions Systems Career Ladder 403X0 - Biomedical Equipment Maintenance Career Ladder
Soft	645X0A - Munitions Inventory Management Shredout 914X1 - Mental Health Work Career Ladder	645X0 - Inventory Management Career Field 732X0 - Personnel Career Ladder

In selecting specialties for the test, all Air Force specialties were classified on a technicality scale as hard (aircraft maintenance skills or experience required), medium, or soft (no aircraft maintenance skills or experience required). They were then cross-classified on a homogeneity scale as either homogeneous (generally fewer than 15 job types, moderated by size of the specialty) or heterogeneous. Test specialties were then chosen on the basis of the recency of their OSRs, as well as currency of their task inventories and STSs. The currency requirement facilitates data replication of the actual jobs and matching to the present Job Proficiency Guide program. The Uniform Airman Record was then searched to find the distribution of members, by skill level, across bases and MAJCOMs. This was done to insure adequate numbers of personnel for the evaluation program.

There are two essential questions to be answered in the evaluation of SPOT. The first is "Can a set of management generated proficiency criteria of a more specific nature than the STS statements guide the field supervisor in selecting tasks for training, improving the quality of training and proficiency level of the force?" This is perhaps the easiest of the issues to conceptualize and the most difficult to evaluate within the dynamics of an operational environment. The potentially confounding variables, such as test environment, experimental factors, practice effect, and motivation, are only somewhat controllable in such field studies.

The second question is whether the occupational survey data base can provide, through prudent choice of item selection screens, task listings which reflect the duties of a particular job position within an AFS. The evaluation of this issue will depend largely on the technical expertise of the supervisors/trainers participating in the study and their ability to express the task performance requirements of the subject position. Any evaluation of SPOT must speak to both of these issues.

Phase 1 of this two-stage evaluation will be a monitored, field supervisor assessment of the item selection screens used to generate SPOT. Each participating supervisor/trainer will initially be presented with a set of SPOTs, one for each job type identified in his AFS. This will range from six SPOTs for the Aircraft Electrical System career field to 27 SPOTs for the Munitions Systems career field. The supervisor/trainer will select the most appropriate SPOT for the subject position and will delete tasks that should not be included and add those on which he requires a proficiency check. The supervisor/trainer will review his modified SPOT lists. The modified SPOT lists will be gathered together with attitude survey and structured interview data to give a criterion for useful SPOT lists. Policy capturing models using occupational survey and other task factor data will be developed in an attempt to replicate the decisions of the supervisors. In addition, data will be gathered from the participants as to the usefulness and administrative burden encompassed in applying SPOT. The results could show that:

a. The experimental screen levels provide SPOT lists which, with only minimal additions and deletions, describe the tasks performed in particular job positions.

b. By raising or lowering the screen levels, useful SPOT lists can be compiled.

c. By using additional analytic techniques, adequate SPOT lists can be generated.

d. The occupational survey data base is appropriate for generating SPOT lists (i.e., few tasks must be added which do not appear in the data base generated lists, and few of the original tasks must be deleted).

The results may well vary across AFSs due to the degree of homogeneity of tasks, differences in OMC data analysis by AFS, and so on. The sampling plan will permit investigation of any variations found. The plan calls for a minimum of 15 supervisors in each specialty to use one of the variants. Because not all MAJCOMs have adequate numbers of personnel at the same bases, six bases will be needed to gather the initial data of screening procedures. The total number of expert supervisors that will provide SPOT assessment will exceed 240 in the test of selection procedures.

Following the staffing of SPOT lists through MAJCOMs, a six month evaluation of the contribution of SPOT to training quality and/or quantity will follow the acceptance of the best selection process from the Phase 1 evaluation. The performance of SPOT trainees on tasks for which they have been qualified will be compared with the performance of a similar group of airmen on these same tasks. Performance will be judged by senior airmen who are subject matter experts in the areas to be judged. Measures will include the number of tasks passed and number failed in proficiency checks. Background data will be gathered in an attempt to identify the effects of confounding variables.

SUMMARY

The SPOT lists compiled by the selection process reported here have an apparent face validity. However, tne objective of SPOT is to provide guidance to field supervisors by standardizing the selection of tasks for inclusion in hands-on OJT. The validity of the selection process has to be established in the Phase 1 evaluation and a best variant of the screening needs to be identified. If this basic utility of SPOT is confirmed, then a second phase of the evaluation will be undertaken to develop an operational SPOT program.

REFERENCES

Ammerman, H. L. Relating Task Surveys to the Content of Existing Training Programs. Paper presented to 19th Annual Conference of the Military Testing Association, US Air Force, San Antonio, Texas, October 1977.

Archer, W. D. Computation of Group Job Descriptions from Occupational Survey Data. PRL-TR-66-12. Lackland AFB, Texas: Personnel Research Laboratory (AFSC), December 1966.

Carpenter, J. B. Sensitivity of Group Job Descriptions to Possible Inaccuracies in Individual Job Descriptions. AFHRL-TR-74-6. Lackland AFB, Texas: Occupational Research Division, Air Force Human Resources Laboratory (AFSC), March 1974.

Christal, R. E. Stability of Consolidated Job Descriptions Based on Task Inventory Survey Information. AFHRL-TR-71-48. Lackland AFB, Texas: Personnel Research Division, Air Force Human Resources Laboratory (AFSC), August 1971.

Stephenson, R. W. and Burkett, J. R. On-the-Job Training in the Air Force: A System Analysis. AFHRL-TR-75-83. Lowry AFB, Colorado: Technical Training Division, Air Force Human Resources Laboratory (AFSC), December 1975.

Watson, W. J. The Similarity of Job Types Reported from Two Independent Analyses of Occupational Data. AFHRL-TR-73-58. Lackland AFB, Texas: Occupational Research Division, Air Force Human Resources Laboratory (AFSC), February 1974.

UTILIZING THE COAST GUARD AUXILIARY IN DEVELOPING INSTRUCTIONAL MATERIALS

Jerrold Markowitz

U. S. Coast Guard Headquarters, Office of Boating Safety, Auxiliary and Education Division, Washington, D. C. 20590.

INTRODUCTION

One of the missions of the Coast Guard is to "minimize loss of life, personal injury and property damage on, over and under the high Seas and Waters subject to U. S. jurisdiction."1 Recreational boating safety is one of the programs within this mission. The purpose of the recreational boating safety program is to reduce the risk of loss of life, personal injury,and property damage associated with the use of recreational boats and to provide boaters with maximum safe use of the nation's waterways".2

The Coast Guard Auxiliary, numbering more than 41,000 men and women, is the official, civilian, volunteer component of the Coast Guard, which supports the boating safety program. The Auxiliary carries out its' responsibilities in 3 general areas: public education, vessel examination, and operations.

In order to increase the utilization of the Auxiliary in operational duties which will supplement regular Coast Guard forces, the quality of Auxiliary training needs to be emphasized. A first step in enhancing quality is in the design and development of instruction. For many years, the Auxiliary has been developing its' own instructional materials and conducting its' own training, monitored by the Coast Guard.

It is important that the Auxiliary continue to be the developers of instructional materials for a variety of fundamental reasons:

1. Extensive local area knowledge of recreational boating;

2. Additional personnel resources at low cost;

3. Individual and Organizational Satisfaction - Developing Instructional materials provides Auxiliarists a unique way of applying their experience. When Auxiliarists learn that other individuals in their group participated in significant ways in the development of a course, acceptance and other positive reactions increases; and,

4. Better communication between students - The person who is experienced in a particular subject can provide better content or substance to a course than a person who has little or no personal experience with the subject. An experienced person can often communicate the real-world point of view, better than the inexperienced person.

[1] The Coast Guardsman's Manual, Sixth edition, U. S. Naval Institute, Annapolis, Md. 1976.

[2] Recreational Boating Safety Operating Program Plan, U. S. Coast Guard FY 82-91.

A new operations Course is currently under development which will train selected Auxiliarists to the level of Coast Guard training, so that the Auxiliarists can supplement the Coast Guard in various search and rescue operations. Practical differences between the Auxiliary and the Coast Guard warrant the development of a new course.

The purpose of this paper is:

1. to discuss a procedure for developing this new operations training course, and

2. to present preliminary results of the development of final examination questions, the first part in the development of instructional materials.

PROCEDURE

Three Auxiliarists were recommended by the Auxiliary Department of Member Training. This Department is one of the Auxiliary National Staff Departments, which is equivalent, organizationally to Coast Guard Headquarters. Each of the three Auxiliarists (who will be referred to as team chiefs) agreed to form teams selected from local flotillas (a flotilla is the fundamental unit of the Auxiliary which is equivalent, organizationally, to field troops). Each of the team chiefs were provided 5 copies of the following items, to be provided to each of their team members:

1. course summary;

2. course objectives;

3. list of course materials that need to be developed; and,

4. guides to assist in development.

Each of these items will be described briefly.

Fig 1 - OUTLINE OF COURSE SUMMARY

I. TITLE
II. JUSTIFICATION
III. COURSE GOAL
IV. COURSE CONTENT

A. Prerequisites
B. Type of Instruction
C. Training Aids
D. Study Material
E. Study Guide
F. Instructor's Guide
G. Written Examination
H. Practical Demonstrations
I. Pre - test

The course summary (fig. 1, above) is a basic overview of the course. It includes:

1. a brief statement of the justification of the course;

2. a description of course goals, specifically what the student will be able to do after completing the course;

3. a brief listing of course content: prerequisties to the course , type of instruction, course materials, pre-and post-test information, and the general subject areas to be covered.

Fig 2 - SAMPLE COURSE OBJECTIVES

T = Theory /Knowledge - written test
P = Skill - Practical Demonstration

SUBJECT AREA	OBJECTIVE	TYPE OF INFORMATION/OUTCOME	ESTIMATED INSTRUCTIONAL TIME
A. FIRST AID	1. Identify the pressure points	T	
	2. Apply commonly used bandages	T/P	
			1 hour
B. COMMUNICATIONS	1. Identify the following radio urgency calls a. MAYDAY b. PAN c. SECURITE d. AUTO ALARM	T	
	2. Identify the necessary Information required to proceed on a distress case	T	
			1 hour

The course objectives (fig. 2, above)were written in behavioral terms and were organized into more specific subject areas. One of the purposes for organizing the course objectives into subject areas was to assign each of the teams different subject areas, in order to develop all the course materials for that subject area. Also included is information on the final outcome for each objective: practical demonstration (for a skill) or a written test (for knowledge), required by the student.

The list of course materials specifies the major products needed, and the order in which they should be developed and submitted to Coast Guard Headquarters, Specifically, they are:

1. Multiple - choice final examination questions

2. Check lists for each skill objective

3. Study Guides

4. Instructor Guide (master lesson plan).

A milestone schedule was also provided.

Fig 3 - GUIDES TO ASSIST IN DEVELOPMENT

1. INTRODUCTION TO WRITING QUESTIONS
2. DEVELOPING FINAL EXAMINATION QUESTIONS
3. INTRODUCTION TO DEVELOPING PRACTICE QUESTIONS
4. DEVELOPING COURSE REVIEW QUESTIONS
5. WRITING MULTIPLE - CHOICE QUESTIONS
6. WRITING MATCHING QUESTIONS
7. WRITING COMPLETION QUESTIONS
8. WRITING TRUE - FALSE QUESTIONS
9. PROVIDING READING OR STUDY MATERIALS
10. DEVELOPING A SHORT SUMMARY OF EACH ASSIGNMENT
11. DEVELOPING AN INTRODUCTION TO A LESSON
12. STUDY GUIDE FORMAT
13. INSTRUCTOR'S GUIDE FORMAT

The guides for developing the course materials (fig. 3, above), includes a brief description on how to develop each of the materials and its' components. The guides are organized in the order in which each of the items should be completed. Also included are samples of completed items. These guides are intended to decrease any confusion, frustration, or loss of time that may arise due to not knowing what is needed or how to do it. An attempt has been made not to include so many details to the extent that they could inhibit a person's creativity or initiative.

The team chiefs were informed that the course materials would be developed using existing Coast Guard and Auxiliary materials that are directly relevant to the course objectives. A logical determination of the appropriateness of the existing materials would have to be made using as a basis the most relevant Coast Guard lesson plans and the course objectives The team chiefs were also informed that if any existing course item is <u>not completely</u> adequate, then that item would have to be modified or completely rewritten. Each of the teams were provided sufficient copies of lesson plans and existing materials.

It was agreed that during the development of this course that one project officer from Coast Guard Headquarters and each of the team chiefs would maintain direct communication without going through the organizational chain. It was agreed that it would be most effective to work as one group. Close liaison between all team chiefs was maintained.

Three specific guides were provided to assist in the development of the final examination questions:

1. Introduction to Writing Questions - This briefly describes the characterstics such as meaningfulness, understandability, conciseness, and preciseness, that all questions should contain.

Fig 4 - FINAL EXAMINATION MATRIX

SUBJECT AREA	NUMBERS OF QUESTIONS PER OBJECTIVE			NUMBER OF QUESTIONS FOR SUBJECT AREA
A. FIRST AID	1 2	2 2		4
B. COMMUNICATIONS	1 4	2 2		6

TOTAL # 10

2. Developing Final Examination Questions - This includes a matrix (fig. 3, above) which estimates the numbers of questions that should be written per subject area (for this course we have decided that the final examination will be 75 questions). The numbers themselves are not to be taken literally, but it is an indication of where the emphasis of the course will be placed, and the emphasis of the final examination. The team chiefs recognized the fact that many more questions would have to be written above the estimated amounts.

3. Writing Multiple - Choice Questions - because we have determined that the final examination will consist of multiple - choice questions, this guide specifies the characteristics of valid questions and includes sample questions. Final examination questions needed to be written for over 100 objectives in 15 subject areas.

RESULTS

The deadline for submitting the first product, the final examination questions, was met. The questions were reviewed by Coast Guard Headquarters. The focus of the review was on whether or not each question is directly relevant to an objective, and if each question is meaningful and understandable.

It has been initially determined that the questions submitted, over 100, are relevant, meaningful and understandable. A more extensive review of the questions is underway in order to put together a final examination for the course.

CONCLUSIONS

The procedure set up appears to be an effective approach that can produce quality products in a short period of time, as indicated by the final examination questions submitted. Comments indicated that participants were well-motivated because the development approach was systematically designed and allowed persons to utilize their technical experience. Specifically:

1. they understood clearly what they needed to do;

2. they were provided guides on how to do it; and,

3. they were provided materials and assistance when needed.

EXTENSION TRAINING MATERIALS: DIFFERENTIAL PERCEPTIONS AMONG SELECTED USAREUR MISSILE PERSONNEL

Raymond O. Waldkoetter and John R. Milligan
US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P. O. Box 33066, Fort Sill, Oklahom 73503

21st Military Testing
Association Conference, 15-19 October 1979
San Diego, California 92152

EXTENSION TRAINING MATERIALS: DIFFERENTIAL PERCEPTIONS AMONG SELECTED USAREUR MISSILE PERSONNEL

Raymond O. Waldkoetter and John R. Milligan[1]
US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P.O. Box 33066, Fort Sill, Oklahoma 73503

INTRODUCTION

The US Army has made a substantial commitment to the reduction of formal MOS producing service school training in an effort to reduce the costs of personnel and training which in many cases may be better performed at the unit level. The reduction in training at MOS producing schools has been effected by the utilization of extension training materials (ETM) which are training materials developed to be used at the unit level by company grade officers and NCOs in training personnel. In most cases this material is intended to be used to keep the individual current and knowledgeable in both his individual military occupation speciality (MOS) and general military subjects common to most MOS's. The use of unit training with ETM in conjunction with supervised on-the-job training is considered sufficient training to award many previously untrained individuals an MOS based upon this type of training. Among the various ETM available for unit training Training Extension Courses (TEC) represent the most comprehensive attempt to provide a block of instruction as a complete audio-visual package to be used by individual soldiers or in small-group settings. Each TEC lesson costs approximately $15,920 (1975 dollars) to develop which compared to conventional group instruction techniques is less expensive Temkin et.al. (1975). This replacement of conventional instruction by TEC was seen as a considerable savings in training costs.

Prior research (McCluskey and Tripp, 1975) demonstrated moderately positive attitudes toward the use of training extension course (TEC) materials by individual soldiers and a more favorable reception to TEC than the traditional lecture method of instruction. Among the observations made by McCluskey and Tripp as a result of their research were the following:

(1) The TEC system had not been implemented in the units for a sufficient period of time to permit a fully adequate evaluation (1975).

(2) During the time period for the evaluation, there was not an adequate library of lessons available for meaningful study in terms of either MOS proficiency or the accomplishment of unit training goals. During the February-March 1975 time frame, when the data was collected only 46 lessons were available to field units.

(3) The results of the evaluation were probably biased by conditions that existed within the 38th Infantry Division since this unit accounted for approximately 80% of the total lesson utilization.

(4) From September 1974 to February 1975, there was an increasing tendency for TEC materials to be utilized in the group mode, during duty hours, and for mandatory study.

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

(5) Command emphasis and information concerning the TEC system were apparently reduced in content and importance during transmission down through the chain of command.

(6) The attitudes of unit trainers and users toward the TEC system were moderatcly positive.

(7) The establishment of battalion-level TEC Learning Centers did not appear to be the most appropriate and effective level for distribution of TEC materials.

(8) During the evaluation, the research team made several observations concerning possible changes that might increase the utilization and efficiency of the TEC system. These observations were supported by relatively small groups (2-10) of personnel interviewed:

(a) Utilization might be increased by thoroughly promoting and demonstrating the TEC Program to Unit Training Officers and NCOs at the company level. This promotion might include a prctotype unit training program that would demonstrate precisely how TEC materials could be applied in a unit environment.

(b) The utilization of the TEC Program in unit training might be increased if the allocation of TEC hardware and software were divided between battalion and company levels.

(c) The utilization of TEC materials for some of the classroom lecture instruction currently given in the units might be increased if the equipment had a projection capability for groups of 30 to 200 individuals.

(d) Utilization of the TEC system might be increased if the TEC Learning Centers had authorization for full-time personnel positions and operating budgets.

(e) Utilization of the TEC system might be increased with a system of rewards and incentives, such as the award of promotion points, for both the student and the unit trainer.

(f) The efficiency of the TEC maintenance system might be increased if three simple operating adjustments were decentralized to the battalion level.

A substantial amount of the data gathered by the authors of this paper five years later (Waldkoetter and Milligan, 1979), essentially supports the recommendations by McCluskey and Tripp (1975) and reflect the failure of the Army to adequately respond to those recommendations. Bennik, Hoyt and Butler (1978) addressed some of these problems by evaluating and suggesting media alternatives for training extension courses in the FY 78-83 time frame. Their findings included:

1. The need exists for: (a) closer attention to the characteristics of soldiers; (b) increased realism of delivery system components; (c) selec-

tion of techniques less demanding of costly resources; (d) closer integration in the choice of training delivery systems.

2. Life cycle management should include integrating system design with: (a) man-machine interface; (b) personnel selection or job assignment criteria; (c) EPMS/OPMS specialty and skill level structure.

3. Choices among the several training delivery systems potentially available in the FY 78-83 period should consider: (a) broadened exportability to include training delivery systems that can be embedded in a fielded weapon system or which can be accessed from a remote site; (b) established data files containing characteristics, operational status, accessibility, and constraints of training delivery systems.

4. TRADOC goals suggest that it is necessary to: (a) insure that course designers developers possess the skills for selecting, developing and updating media and courseware for a variety of alternative delivery systems; (b) ensure that school system managers can specify procurement requirements as well as monitor and evaluate contractor plans and products; (c) collect and summarize data on training cost effectiveness to include user acceptance throughout the life cycle development of a system.

Knerr, Downey, and Kessler (1975) compared effectiveness of TEC to conventional instruction in a field experiment of both Active (N=635) and Reserve (N=539) Army Components. Their results demonstrated that TEC trained soldiers performed significantly higher on post tests than conventionally trained soldiers. A finding of interest in the research was that performance test scores were uncorrelated with General Technical (GT) aptitude scores for soldiers trained using TEC materials while soldiers trained by conventional instruction obtained performance scores which were correlated with their GT scores. This suggests the possiblity that TEC can be used with a wide range of individuals with differing levels of preparedness for study. Among the most recent research is a study by Holmgren, Killigross, Swezey and Eakings (1979). This research used four experimental groups and a control group to measure training effectiveness and retention of training extension course materials in five subject areas of a pre-post experimental design. The results found again that soldiers trained using TEC performed better than soldiers trained through conventional methods.

The research cited above is representative of research in the area of exportable or extension training materials. Much of this research has focused upon TEC materials and has not addressed the other areas of extension training methods or materials. Although several studies have assessed usage and availability of materials there have been no follow-up studies to evaluate whether the recommendations made in earlier studies have been implemented or to evaluate whether the materials have been used with increasing frequency as the availability of materials increase at the unit level. Research may demonstrate as suggested here that extension training materials most often are effective in training but without their utilization by the individual soldiers and trainers their value comes into question.

The research conducted by the authors in this study sought both qualitative and quantitative data on extension training at the unit level. Lance units located in Europe were selected due to their relative isolation and importance to the defense of Western Europe.

METHOD

The research reported here attempted to provide answers to three major areas of doubt regarding extension training materials and its use by Lance missile units. These areas of concern were (1) actual usage of materials and by whom (2) availability and suitability of materials including the training environment itself and (3) individual perceptions as to quality and desirability of the extension training program. It was hoped that meaningful answers to these three areas of inquiry would provide a basis for the re-allocation of training resources to more fully meet the needs of both the individual soldier and management needs of the Field Artillery School, the proponent agency for Field Artillery extension courses.

To accomplish the above research objectives the authors used both written surveys and structured interviews of personnel (N=323) in US Army Lance units located in West Germany. This technique proved to be effective and efficient in accomplishing the research objective.

Sample. The researchers coordinated data gathering with the headquarters of six US Army Lance missile units in the Federal Republic of West Germany. The data gathering was arranged so that all available soldiers of the surveyed units would complete the questionnaire in conjunction with scheduled unit training. Surveyed units were asked to not change any personnel or training schedules but simply provide a minimum number of unselected personnel to complete the survey. Although formal random sampling techniques were not used, the researchers were satisfied that no systematic bias in subject sampling was present. Analysis of the collected data with regards to rank, MOS and prior research confirmed the researchers observations. Of the individuals present less tha 10 refused to complete the survey although a total of 30 questionnaires were excluded from the analysis due to incompleteness of responses leaving a sample of 323.

Research Instruments. The questionnaire used in this study was developed by the researchers in conjunction with representatives from the Field Artillery School. It is a factorially complex instrument whose psychometric characteristics will be reported elsewhere. The purpose of this report is to report descriptive responses to selected individual items rather than an analysis of the instrument itself. Copies of the complete questionnaire results are available from the researchers and are not presented here due to space limitations.

RESULTS

Inspection of the questionnaire responses clearly indicate that the most available materials are field manuals (96%), technical manuals (95%) and TEC programs (85%). Frequency of use indicated that technical manuals (42.2% very often) lead field manuals (38% very often) in usage with TEC programs being given a frequency of use of only 2% (very often), 11% (often) and 40% (occasionally) with the most infrequent use being made of television or closed circuit T.V. (73% never). Responses to the question of which media materials the respondents found most helpful in learning indicated that formal technical manuals were rated 37% (extremely helpful) and field manuals 30% (extremely helpful) with TEC programs rated at 12% (extremely helpful).

Respondents to the question of quality of the extension materials provided the unit, felt again, that technical and field/soldiers manuals were of the highest quality in comparison to other media provided. Respondents to the question of the importance of various training materials in helping the individual learn and retain proficiency in his MOS reflected again the importance of field and technical manuals over other forms of information including TEC.

Comparison of the percentage of responses by rank are shown on Tables 1 and 2. These responses by rank of respondent (enlisted, NCO and Officer) provide evidence as to differing perceptions of the utility of two selected media modes (TEC and TMs). Officers stated TEC was more available than that stated by the NCOs who in turn stated TEC was more available than the enlisted personnel. These differences between groups for the five selected questions dealing with TEC were all statistically significant beyond the .05 level suggesting substantial disagreement on the five TEC dimensions of availability, frequency of usage, job learning aid, quality, and importance. Of particular concern is the observation that 60% of the enlisted persons use TEC only infrequently or never.

Comparison of percent response by rank for the five questions dealing with technical manuals (TM) reveals statistically significant differences among groups on four of the five questions (substantial agreement among groups existed on the availability question) but unlike the TEC differences the overall responses for all ranks rated TM consistently higher on the dimensions of availability, usage, learning aid, quality and importance. This strongly suggests that TEC has a long way to go before it becomes as significant a job aid as TMs are currently to soldiers at unit level.

DISCUSSION

The questionnaire and interview results strongly support the findings of McCluskey and Tripp (1975) with the observation that those findings are still current for Lance missile units in West Germany with the recommendations in that report having not been implemented five years later. A major justification after implementing TEC and other forms of ETM has been its projected cost-effectiveness in replacing much of the conventional instruction at unit level and providing a supplement to MOS training. What the researchers have found is that TEC is not replacing conventional instruction but in some instances does serve as a little-used supplement to conventional instruction. It appears to the researchers in this study that TEC has added substantial costs to unit training rather than reducing those costs and prior cost-effectiveness analyses are not currently accurate due to the failure to implement the McCluskey and Tripp recommendations.

Table 1

Comparison of Percent Responses by Rank to Training Extension Course (TEC) Questions

Availability of TEC

	Enlisted Persons	NCO	Officer
Yes	79.2	92.0	96.2
No	17.6	8.0	3.8
No Answer	3.1	0	0
Sample =	159	75	52
Missing Observations =	37		
Chi Square =	13.88	4d.f.	p=.01

Frequency of Using TEC in Training

	Enlisted Persons	NCO	Officer
Never	35.6	11.4	7.5
Infrequently	24.3	17.7	26.4
Occasionally	35.6	50.6	39.6
Often	4.5	19.0	18.9
Very Often	0	1.3	7.5
Sample Size =	177	79	53
Missing Observations =	14		
Chi Square =	54.17	8d.f.	p=.001

How Helpful is TEC in Your Learning

	Enlisted Persons	NCO	Officer
No help	14.9	9.0	2.0
Somewhat helpful	22.6	16.7	21.6
Helpful	37.5	33.3	43.1
Very helpful	12.5	30.8	19.6
Extremely helpful	12.5	10.3	13.7
Sample Size =	168	78	51
Missing Observations =	26		
Chi Square =	18.02	8d.f.	p=.02

Table 1 (Continued)

What is the Quality of TEC Provided Your Unit

	Enlisted Persons	NCO	Officer
Very poor	15.6	16.0	4.1
Poor	13.0	18.7	28.6
Satisfactory	39.0	24.0	20.4
Good	24.7	29.3	26.5
Excellent	7.8	12.0	20.4
Sample Size =	154	75	49
Missing Observations =	45		
Chi Square =	21.09	8d.f.	p=.01

Importance of TEC in Learning Your Job

	Enlisted Persons	NCO	Officer
Not important	9.7	3.9	0
Slightly important	15.8	19.5	33.3
Moderately important	34.5	22.1	33.3
Very important	26.1	37.7	23.5
Extremely important	13.9	16.9	9.8
Sample Size =	165	77	51
Missing Observations =	30		
Chi Square =	19.68	8d.f.	p=.01

Table 2

Comparison of Percent Responses by Rank to Technical Manual (TM) Questions

Availability of TM

	Enlisted Persons	NCO	Officer
Yes	95.2	94.8	96.2
No	3.6	5.2	3.8
No Answer	1.2	0	0
Sample Size =	165	77	53
Missing Observations =	28		
Chi Square =	1.91	4d.f.	p=.75

Frequency of Using TMs in Training

	Enlisted Persons	NCO	Officer
Never	7.6	5.0	0
Infrequently	8.8	3.8	3.9
Occasionally	20.0	10.0	9.8
Often	32.4	30.0	21.6
Very Often	31.2	51.2	64.7
Sample Size =	170	80	51
Missing Observations =	22		
Chi Square =	25.85	8d.f.	p=.01

How Helpful are TMs in Your Learning

	Enlisted Persons	NCO	Officers
No help	1.8	2.6	0
Somewhat helped	15.2	1.3	5.9
Helpful	23.8	21.1	23.5
Very helpful	27.4	28.9	29.4
Extremely helpful	31.7	46.1	41.2
Sample Size =	164	76	51
Missing Observations =	32		
Chi Square =	15.81	8d.f.	p=.05

Table 2 (Continued)

Quality of TMs Provided to Your Unit

	Enlisted Persons	NCO	Officers
Very poor	7.0	2.8	1.9
Poor	6.3	15.3	17.3
Satisfactory	18.4	18.1	19.2
Good	36.1	30.6	44.2
Excellent	32.3	33.3	17.3
Sample Size =	158	72	52
Missing Observations =	41		
Chi Square =	14.26	8d.f.	p=.08

Importance of TMs in Learning Your Job

	Enlisted Persons	NCO	Officers
Not important	2.4	0	0
Slightly important	9.6	6.4	17.6
Moderately important	18.0	11.5	7.8
Very important	31.1	30.8	31.4
Extremely important	38.9	51.3	53.1
Sample Size =	167	78	51
Missing Observations =	27		
Chi Square =	12.35	8d.f.	p=.14

REFERENCES

Bennik, F. D., Hoyt, W. G., and Butler, A. K. Determining TEC media alternatives for field artillery individual-collective training in FY 78-83 period. ARI Report TR-78-A3 (ADA 053528). Army Research Institute for the Behavioral Social Sciences, Alexandria, Virginia, February, 1978.

Holmgren, J. E., Hilligross, R. E., Swezey, R. W., and Eakings, R. C. Training effectiveness and retention of training extension course (TEC) instruction in the combat arms. Research Report 1208, Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia, April 1979.

Knerr, C. S., Downey, R. G., and Kessler, J. J. Training individuals in Army units: Comparative effectiveness of selected TEC lessons and conventional methods. Research Report 1188 (ADA 022034), Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia, December 1975.

McCluskey, M. R., and Tripp, J. M. An evaluation of the utilization, maintenance, and perceived benefits of the training course (TEC). Human Resources Research Organization, Alexandria, Virginia, June 1975.

Temkin, S., Connolly, J. A., Marvin, M. D., Valdes, A. L., and Caviness, J. A. A cost assessment of Army training alternatives. Research Problem Review 75-3, Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia, August 1975.

IMPLEMENTATION OF EXAMINATIONS INTO THE AIR WAR COLLEGE RESIDENT PROGRAM

by

Brian K. Waters; LaVerne G. Junkmann; William R. Vlach

Air War College
Maxwell AFB, Alabama

Paper presented at the Military Testing Association
Annual Meeting, San Diego, California, 16 October 1979

IMPLEMENTATION OF EXAMINATIONS INTO THE AIR WAR COLLEGE RESIDENT PROGRAM[1]

Brian K. Waters; LaVerne G. Junkmann; and William R. Vlach[2]

Air War College
Maxwell Air Force Base, Alabama 36112

Introduction

The Air War College (AWC) is the Air Force's senior professional military school. Its mission is to prepare selected officers for key command and staff assignments where they will be tasked with responsibility for developing, managing, and employing airpower as a component of national security. The curriculum is broken into three major courses: Leadership and Management, National Security Affairs, and Military Employment. Students spend approximately 48% (about 19 hours per week) of their time in class with the bulk of the rest of the hours devoted to research and independent study.

The environment of the college is oriented toward a free expression of ideas and an opportunity for independent, analytical, and creative thinking. Diverse military doctrines, policies, and strategies are examined. Students and faculty members examine current and future Department of Defense problems with a view toward providing solutions.

The AWC resident program has one 10-month class per year with an AY79 quota of 228 students. Each of the military services is represented along with selected civilian agencies of the Federal Government and officers from selected foreign countries. A voluntary nonresident program consisting of seminar and correspondence modes is conducted for senior officers not selected for the resident program. This paper will focus on the AWC resident program for academic year 1978-79.

Student Population:
Air Force AWC students represent about the top 15% of officers in the 16-21 year group. Table 1 provides descriptive statistics on the academic year 1978-79 (AY79) class. Two-thirds held graduate degrees, and as a group, represent the top of their profession.

Table 1
AY79 CLASS DESCRIPTIVE STATISTICS

Agency	
US Air Force	170
US Army	22
US Navy	5
US Coast Guard	1
US Marine Corps	6
Foreign Officers	14
Civilians	10
	228

Highest Education (US)	
High School Diploma	8 (4%)
Bachelors Degree	58 (22%
Masters Degree	132 (66%)
Doctorate Degree	18 (8%)

Median Age = 40 years
Median Years of Service = 17 years

[1] This paper reflects the academic year 1978/79 AWC experience and philosophy. Changes in the AY 1980 curriculum and testing procedures are not included.

[2] Views or opinions expressed or implied are of the authors and are not to be construed as carrying official sanction of the Air University (ATC) or the Department of the Air Force.

AWC Faculty.

Ninety-five percent of the AWC faculty of fifty-eight hold graduate degrees with twenty-eight percent PhDs or equivalent. They represent all four services, two foreign nations, and include nine civilian full-time faculty members.

With this brief description of AWC, its faculty and students, we now turn to the main purpose of this paper: to describe AWC's experience in implementing an essay examination program into the resident curriculum. The balance of the paper will answer four main questions: (1) Why test?; (2) How did we do it?; (3) How well did it work?; and (4) What did we learn?

Why Test?

Historically, testing in education has served two main purposes: to evaluate student performance and to evaluate curriculum effectiveness. From AWC's perspective, the primary reasons for examining individual student achievement should be to improve instruction and to facilitate increased learning. The grades that result from examinations should be a secondary objective.

The use of properly prepared and administered examinations can have an immediate and positive effect on the learning achievement of students. Properly constructed examinations that test valid, non-trivial, learning objectives can motivate students to attain objectives by stimulating increased learner activity, by directing attention to the desired learning outcomes, and by providing feedback that enables the student to assess personal weaknesses.

Student transfer of learning and retention are also facilitated by examination of learning achievement. This is particularly true of examinations that measure the more complex, higher levels of learning. Reinforcing practice of such complex activity as analyzing multi-faceted problems, formulating plans, and making evaluations and decisions under uncertainty and risk increase the students' retention of principles and concepts and provide exercise in their application to new situations.

In addition to the influence examinations can have on student achievement, they also provide a means for improved curriculum and instructional methodology decisions. The results of examinations can provide the faculty with an effective means of determining the degree to which course content has been understood. Test results can guide the selection of appropriate course content and instructional methodology to achieve optimal student learning.

If examinations are used to improve the instructional program and increase learning and are not used to threaten or rank the students, they might be more likely to accept testing as evidence of learning progress and as an aid to the identification of areas needing further improvement.

AWC Instructional Objectives.

The instructional objectives of the AWC emphasize professional education rather than training as a central purpose. The objectives seek to develop understanding rather than the memorization of principles and procedures. The dominant aim is to develop insight, breadth of vision, and a capacity for dispassionate analysis. The curriculum is designed to raise the professional stature of the students by providing them with a conceptual framework and the intellectual tools which will enable them to understand, analyze, and critically evaluate the role, capabilities, and limitations of airpower.

Based upon this philosophy of AWC education, the decision to begin the examination of individual student achievement in the AWC was announced to the staff and faculty by General Schoeneman, AWC Commandant, who stated that:

> ... we need to know how well each student has learned what we set out to teach. A testing program will provide a feedback to the student, information on his progress, and a measure of program effectiveness.

Thus, AWC was committed to introduce examinations into our program. The next question was how.

How Did We Do It?

Types of Examinations.

Examinations at the AWC should be appropriate for measuring achievement of AWC instructional objectives. As discussed in the previous section, the AWC objectives focus on the complex behaviors of critical analysis, formulation (synthesis), and logical evaluation.

The remainder of this section will consider the two basic types of tests--selection and supply--and their relative merits for use in assessing student achievement of AWC instructional objectives. The selection, or objective, type of test offers the students two or more alternatives from which they select a single best answer. The multiple-choice question is the most common example of the objective or selection type of test item. Matching and true-false questions are other examples. The second type of question is the supply type. The essay question is an example of this type of question. Table 2, adapted from Gronlund, shows a comparison of the two types of tests.

Table 2
COMPARISON OF OBJECTIVE AND ESSAY TESTS

Measure of Relative Merit	Objective Tests	Essay Tests
Level of learning measured	Good for lower levels of learning. Inadequate for synthesis and evaluation.	Good for higher levels of learning. Best for synthesis and evaluation.
Preparation of test items	Preparation of good items is difficult and time consuming.	Preparation of good items is difficult, but easier than objective type.
Scoring (grading)	Objective, simple, and highly reliable.	Subjective, difficult, and less reliable.
Probable Effect on Learning	Encourages students to remember, interpret, and analyze the ideas of others.	Encourages students to organize, integrate and express their own ideas.

Gronlund, Norman E., Constructing Achievement Tests, New Jersey: Prentice Hall, Inc., 1968, p. 68.

Considering that the dominant aim of the AWC instructional objectives is to develop insight, breadth of vision, and a capacity for dispassionate analysis leading to the formulation of plans (synthesis) and effective decision-making (evaluation), the most appropriate type of test would be the essay as indicated by the first measure of merit shown in the table above. The second measure in Table 3 concerns the preparation of test items. While difficult, the essay test item is easier to prepare than the objective type of test item.

While essay examinations are more difficult to score and the grades tend to be less reliable, scoring difficulty can be lessened by the preparation of adequate evaluation criteria prior to grading. Further, the lesser reliability of essay examination grades is of less significance in the AWC than it might be in other institutions since the major purpose of the examination is improved learning rather than discrimination or ranking of students.

Finally, the probable effect that essay tests have on learning, i.e., to encourage students to organize, integrate, and express their own ideas, is of major significance in making a choice between types of test items. The graduates of a senior service school and candidates for key command and staff duties must surely be capable of independent analysis and evaluation; they must certainly be able to organize, integrate, and express their own ideas rather than merely select from among the ideas of others. Thus, the decision was made to use essay examinations as the testing mode at the AWC.

Faculty Development.

The addition of examinations to AWC had direct implications on faculty selection and training. Faculty understanding of the matching of curriculum and testing to the levels of learning stated in objectives became very important. Workshops were held by AWC and Academic Instructor School faculty members on Bloom's Taxonomy, instructional systems development (ISD), evaluation, and essay item writing and grading. The initial faculty acceptance of testing into AWC was far from unanimous, with the majority probably negative toward examinations at AWC. It was most important that the faculty understood the rationale for the program prior to implementation.

Building a Pool of Examination Questions.

For each course, a pool of essay questions was developed at the same time that the instructional objectives, readings, and instructional methodologies were selected. During this time the course designers were in the best position to specify how to measure student achievement of the objectives. The grading criteria for each question were also developed at that time, covering the course material presented in each of the 3 courses: Leadership and Management, National Security Affairs, and Military Employment.

The questions that were prepared for the test pool were forwarded by the course chief to the Test Review Committee (TRC) for review and in turn to the Dean of the Resident Program for approval. This review and approval cycle was concurrent with the review and approval of course and block instructional objectives. The TRC was composed of the educational advisor to the Commandant (chair), and representatives from the directorate of evaluation and the curriculum planning division. The TRC was charged with evaluating test quality, specifically looking at content/test item match, test item clarity, administrative procedures, and gradability of the test items. Course curriculum developers were required to supply "model" item responses to the TRC. Since students were given a limit of 2-3 pages in which to answer the items, the scope of submitted questions was also of major TRC concern. "Content" was not implicitly evaluated by the TRC. Once the TRC completed its test item review, a pool of examination questions was provided to the students.

Printing and Distribution of a Pool of Examination Questions.

A pool of essay questions covering material presented in each course was distributed to the students prior to each block of instruction. The students were told that the majority of the questions on their examinations would consist of questions that were identical or nearly identical to those in the question pool; and further, the balance of the questions would be of a similar nature and scope. This strategy told the students what was expected of them and the depth with which they had to deal with the subject material. It was hoped that this would substantially reduce the threat of the examination and increase successful implementation of the testing program.

Selecting Questions for Actual Examinations.

After TRC review, the course chief selected and modified the questions to be used on the actual examination. The course chief was also responsible for the construction of those additional questions that were to be added to the examination. The examinations were closed-book, with a time limit that was considered ample for the preparation of an adequate response.

Grading Procedures for Examination.

To insure a fair and uniform evaluation of each student's responses; a team of faculty members was formed to grade all student responses to a given question. A 13 point grading scale was used running from F to A+ with general guidelines as shown in Table 3 given to graders.

Table 3
GENERAL GRADING STANDARDIZATION GUIDE

+ A -	Factual and fully supported: Complete answer.
+ B -	Factual and little or no support: Relatively complete answer.
+ C -	Minor misconceptions and/or minor gaps in information.
D/F	Serious misconceptions and/or serious gaps in information.

The assignment of +s and -s will be made within grade categories through relative comparison within grade category.

Teams compare scales after standardizing grades of 10 selected papers. Determine scaling equivalence with criterion that all papers should be in same grade category. When criterion achieved through discussion, grade rest of papers and randomly check reliability when finished to assure no change of standards.

In addition, each grader on each team had copies of the "model" answers, amplified by relevant points from the graders as well as actual answers from students. Each team blind-graded (no examinee identification on papers) 225 2-3 page responses on a given question. To begin grader standardization, 10 student responses were randomly selected from the answers to a question. Each of the 2 or 3 graders graded the 10 selected answers independently. Inter-rater reliability was achieved by raters discussing grading criteria on all items that had 2 scale point (e.g., B+ vs A, etc) difference between graders. Once this criterion was met on all 10 papers, the team members graded the other papers separately, with occasional consistency checks done to assure no change of rater standards during the grading process. Once the grading on each examination was completed, evaluation directorate personnel entered grade data into the computer and verified the data base. In all 5 examinations, the exams were graded, returned to students, and the frequency distributions of grades by exam item posted within 5 working days or less of the examination. Expeditious return of the tests to the students was considered a major goal for increasing the instructional value of the examinations.

One other feature in the AWC examination grade procedure was that each question was graded separately with no overall test grades given. This procedure focused student attention on learning as opposed to grades. In addition, extensive written feedback was made by graders on each response to highlight weaknesses or strengths of each answer. Throughout the entire process, the emphasis was on the examination as a learning methodology rather than as a measurement device. As will be seen in the next section of this paper, AWC was not too successful in selling this testing rationale to our students.

How Well Did it Work?

An evaluation of the AY79 examination program must consider several different kinds and sources of measures. This paper will look at 3 primary sources of data: (1) test and item analyses, (2) student attitudinal and observed data, and (3) faculty subjective and objective perceptions. In each source, both positive and negative effects evolved. This section will highlight both, and try to put it all together into an evaluation of the net result of the implementation of the examination program into the AWC.

Test and Item Analysis.

Table 4 shows the structure of the AY79 exams. The Leadership and Management course (Course I) elected to give 3 exams, each covering a phase of instruction, while the other 2 courses each decided to give a single 3-hour comprehensive final examination. Students were given a choice of questions on all 5 tests to reduce student anxiety and to downplay the measurement aspects of the testing situation.

Table 4
STRUCTURE OF AY 79 AWC EXAMINATIONS

Course	# of Items/Exam	Time/Exam
Leadership & Management	Ph 1 - Choose 3 of 4	45 min
	Ph 2 - Choose 3 of 4	60 min
	Ph 3 - Choose 3 of 4	60 min
National Security Affairs	Choose 4 of 5	180 min
Military Employment	Choose 4 of 5	180 min

Table 5 depicts descriptive statistics for the 22 test items and 5 course exams given over the entire year. Several conclusions are immediately evident from Table 5. First, a great deal of grade inflation apparently existed, and concomitantly, little discrimination between student performance in general. These problems were most evident in the Military Employment course. Secondly, the National Security exam was "best" from a measurement standpoint, with less inflation and relatively high reliability and validity. Clearly, the greater test score variance helped, but test content analysis suggested that the items also solicited higher levels of learning than the other courses. One other main element differentiated the course testing--time. In the Leadership and Management course, one hour examinations were given. The first exam required the students to answer 3 questions in 45 minutes; the last 2 Course I exams were each 60 minutes. Test analysis showed that the time pressures in the first test were excessive. Note the extremely low validity of the last item in the first test.

Table 5
DESCRIPTIVE ITEM/TEST STATISTICS FOR AWC AY 79 EXAMINATIONS

	Mean	S.D.	Validity*	Reliability**
Course I - Leadership & Management				
Phase 1 Test	30.39	5.91	.38	.59
Item 1	11.1	1.95	.22	
Item 2	10.4	2.01	.16	
Item 3	10.8	1.94	.37	
Item 4	10.2	2.82	.04	
Phase 2 Test	32.06	5.57	.50	.52
Item 1	10.5	2.12	.35	
Item 2	10.3	2.10	.22	
Item 3	8.9	2.59	.29	
Item 4	9.5	1.86	.19	
Phase 3 Test	29.51	6.15	.60	.68
Item 1	10.5	1.40	.29	
Item 2	9.6	2.29	.44	
Item 3	10.1	2.24	.33	
Item 4	9.1	2.52	.38	
Course II - National Security Affairs				
Final Exam	37.57	7.35	.58	.77
Item 1	9.8	2.10	.50	
Item 2	9.5	1.99	.45	
Item 3	8.9	2.06	.48	
Item 4	9.6	2.23	.56	
Item 5	9.3	1.80	.52	
Course III - Military Employment				
Final Exam	40.71	5.20	.43	.60
Item 1	9.8	2.01	.49	
Item 2	11.0	1.23	.38	
Item 3	10.5	1.12	.31	
Item 4	10.4	1.89	.37	
Item 5	9.3	1.92	.29	
OVERALL (Criterion)	201.6	22.41		

Grade Scale (A+=13, A-=12...D-=2, F=1)

* Correlation with sum of all other measures (other item grades, research paper grades, elective course grades, and written paper grades).

** Hoyt Internal Consistency Reliability Estimate

Examination Program Manpower Costs.

Table 6 shows the estimated manhours used for the AWC AY79 examination program. The total requirement turned out to be about .7 faculty manyears and 12.4 student manyears. Obviously, these figures have to be evaluated considering the opportunity costs involved with what could have been done with this time if it had been used by faculty and students in other efforts. This issue is a major aspect of the arguments of the proponents and detractors of having examinations in the AWC. The basic question comes down to whether the examinations and the time required to implement them were productive.

Table 6
AWC AY79 EXAMINATION PROGRAM MANPOWER ACCOUNTING

Department Personnel	Estimated Manhours
Test Pool Item Development	88
Sample Answer Preparation	20
Test Review Committee Attendance	30
Preparation of Final Test Booklets	70
Grader Standardization	3
Department Chief Grade Review	6
Exam Grading	
Course I	318
Course II	182
Course III	160
Evaluation Personnel	
Exam Data Compilation	40
Item Analysis/Test Analysis	75
Test Report Preparation	25
Test Review Committee (5 members)	
Pre-meeting Preparation	65
Meeting Attendance	165
Supervisory Review	15
Clerical/Logistical Support	35
FACULTY & STAFF TOTAL (.71 Manyears)	1297
Student TestingTime (227 students)	2040
Estimated Student Exam Study Time (10 hours/student/hour testing)	20400
STUDENT TOTAL (12.44 Manyears)	22440
AWC TOTAL	23,737

Faculty and Student Attitudes.

A major concern of the Institution was how faculty and students in a senior service school would accept the examination program.

Students. Student attitudes toward being tested were predictably negative. The students began the year feeling that tests were not appropriate for a senior service school, that they diverted student attention away from more important, non-testable learning outcomes of the college, and that the instution was going to use the grades to differentiate between students rather than just for curriculum evaluation and as a feedback mechanism. Negative attitudes were very strong early in the academic year, but lessened considerably as the year progressed. By the last exam, the students seemed to consider the exams as a nuisance hurdle which they had to face, but they no longer actively resisted the program.

Student study behavior for the exams was fascinating. They teamed up to "game" the tests; they had extensive study group meetings with sample questions and answers distributed, discussed, and shared between study groups. "Laundry lists" of model answers to test pool or suspected questions were evident as exam time approached, and the test pool questions which actually were selected for the exam produced "laundry list" answers in many cases, particularly in Course I. More than any class in memory, students became very critical of learning objectives that were not, from their perception, adequately covered in the instruction. They also became very critical of entertaining speakers who were relatively light on the content objectives. In general, from the faculty members' perceptions, the AY 79 class was more conscious of AWC's instructional objectives than any class in their memory.

At the end of the academic year, AWC surveyed all resident students through a 79-item questionnaire. Two items on the AY 79 survey solicited student perceptions of the exam program. Table 7 shows the items and proportion selecting each response. Clearly, even by the end of the year, the students were not very positive about being tested.

Table 7
ITEMS ON END-OF-YEAR SURVEY

The use of essay examinations was a valid method of providing feedback on what I learned at AWC.	Strongly Disagree	27%
	Disagree	29%
	Undecided/Uncertain	18%
	Agree	22%
	Strongly Agree	4%
The examination program motivated me to apply myself more diligently than if there had been no examinations.	Strongly Disagree	34%
	Disagree	28%
	Undecided/Uncertain	10%
	Agree	22%
	Strongly Agree	7%

Faculty. Faculty attitudes toward examinations, in general, seemed to drastically change during the course of the school year. At the beginning of the year, about three-fourths of the resident faculty would have preferred to forget the whole thing. Their reasons tracked closely with the student rationale. By the end of the year, however, the proportion of faculty examination program supporters was probably 75%, with few faculty members strongly opposed. The feeling among the faculty was that the AY 79 class was working harder, learning more, and challenging the college to improve the instruction more than they had ever seen. The exams were also challenging the faculty members to strengthen their individual weak areas as the students pushed for clearer answers to their questions on the objectives. Faculty members were quick to "see how their guys did" after each exam and individually analyzed areas that appeared generally weaker in their seminars. Resistance to the extensive exam grading time was felt in all three courses, but grading time per item decreased considerably after the first 2 exams and experience became a guide. A final faculty change probably resulting indirectly from the examination program was a greatly increased quality level and frequency of faculty workshops throughout the year. Overall, the faculty members have accepted the examinations into the AWC curriculum as a beneficial aspect of our program.

Overall Assessment.

The use of examinations at an institution like the AWC has both positive and negative effects on the college. Table 8 summarizes the authors' observed pros and cons.

Table 8
PROS AND CONS ON THE EFFECTS OF EXAMS (AY79)

PROS

- o Student and faculty attention strongly focused on instructional objectives.
- o Student comments criticized speakers who deviated from objectives.
- o Heavy student cooperation within and across seminars preparing for exams.
- o Strong pressures on faculty to strengthen individual weak areas.
- o Individual feedback to students on performance.
- o Feedback to faculty on instructional success.
- o Credibility to outside agencies/visitors.

CONS

- o Student negative reaction.
- o Increased faculty workload.

Weighing the quantitative and qualitative costs against the predominantly qualitative benefits makes for a difficult evaluation. The authors feel that overall the evidence is strongly supportive of the examination program, primarily because it focused student and faculty attention on the instructional objectives of the institution. If the instructional objectives validly reflected what AWC graduates need to know to fill "key command and staff assignments," and the faculty strived diligently to make that so, then a program that led to students learning more about the objectives was beneficial to the college. In our opinion, the examination program at AWC fulfilled this goal sufficiently to recommend retention of examinations in the AWC.

What Did We Learn?

AWC was brand new at this large scale essay examination business in AY 79. The authors feel that many of our successes and failures may serve as lessons learned for other institutions considering a similar effort in the future. As with the rest of this paper, the recommendations fall into two major areas--examination development, grading, and analysis; and attitudinal considerations for students and faculty.

Examination Item Writing.

Items should be prepared as the instruction is developed, not just prior to test time. This assures content coverage in the curriculum and that examination items measure important outcomes.

Faculty examination workshops should be held prior to implementation to emphasize item writing techniques.

Examination Grading.

Evaluation personnel should extensively plan and control exam grading. AWC could have done much more in this area than was done last year. Careful training of graders and inter-grader reliability checks during grading would have improved last year's test statistics significantly.

A record should be kept of grades assigned by each grader to help control inflated grades. A major effort should be made to encourage item score variance across students. This has the effect of rewarding outstanding student performance and "getting the attention" or poorer performers. If nearly all students get "A's" a secondary benefit of having the exams is lost.

Examination Length.

Enough time should be allowed to enable all students to complete the exam. AWC would recommend at least 20 minutes per item for 2-3 page essay responses, and preferably at least 30 minutes. Essay items designed to measure synthesis and evaluation levels of learning should have longer page limits and even more time allowed for completion.

Examination Content.

Items should be challenging, thought-provoking questions which require a decision from the student along with support for his/her position. Do not ask specific knowledge level items that encourage "laundry lists." Ideally, there are no right answers to senior service school level questions, only logical, well-supported arguments.

References in questions to particular authors or sources and the use of jargon should be avoided. The objectives of graduate-level testing should not encourage memorization of names, theories, or formulas.

Items should ideally cover more than one instructional objective, making the student synthesize concepts across blocks of instruction.

Sample Test Question Distribution.

Guaranteeing that 50% or more of the examination questions would come from the sample pool distributed to the students was dysfunctional. It encouraged the students to try to game and memorize model answers.

Student Attitudes.

Once the decision is made to test, the institution must make it clear to the students that the decision has been made and is not up for negotiation. Early in the academic year, AWC suffered from much student discontent, as the students appeared to feel that if they "fussed" enough the institution might change its mind about giving examinations.

AWC students did not like being asked knowledge level questions. Many student comments stated that if we had to test, they'd prefer to be challenged with a tough test: reliably and validly graded.

Don't expect students to like being given exams--they won't.

Faculty Attitudes.

Provide extensive faculty guidance on why the testing is being implemented and gain faculty support in developing the program.

Conclusion

Overall, AWC's first venture into testing must be considered a success, though a qualified one. In many ways, we did much better than many of us expected. Much of the credit for that success came as a direct result of the extensive planning which was done in the 8 months prior to the AY79 start. Despite the planning, we still made some mistakes which hurt our program. Our goal in writing this paper has been to help other institutions taking on a similar task to avoid some of our errors and to benefit from some of the things we did that worked. We encourage telephone or written contact by those interested in greater detail about our essay testing program or those with relevant experience that could help us as we enter our second year of AWC testing.

INCENTIVE MANAGEMENT TRAINING:
USE OF BEHAVIORAL PRINCIPLES FOR PRODUCTIVITY ENHANCEMENT

Ste.en L. Dockstader
Delbert M. Nebeker
Jacqueline Nocella
E. Chandler Shumate

Navy Personnel Research and Development Center
San Diego, California 92152

The behavioral approach to personnel management has received increased attention in both research and applications during the past decade (Luthans & Kreitner, 1975; Miller, 1978). The success of feedback, reward, and other incentive management programs can often be attributed to the effective application of known behavioral principles. Such strict application of the principles associated with feedback (specificity, timeliness, frequency, etc.) and rewards (contingency, equity, timeliness, etc.) is not typical of the management of organizations. In fact, the salience of these principles was developed in controlled laboratory settings in psychological experiments designed to examining learning behavior (Marx, 1969). However, the management community "has virtually ignored the findings of empirical psychology, which has been built on the same technological methods that have produced the greatest gains in productivity in the material area (Miller, 1978, p. 2)."

The reluctance on the part of managers to use incentive management is probably from at least two sources. First, the effective application of the principles requires reliable performance measurement that validly reflects work behavior. Highly complex jobs, i.e., those having many and assorted separate tasks, and those jobs which are highly cognitive in nature represent a significant (though not impossible) challenge to such measurement. Second, the behavioral principles have such a common-sensical nature that they are seldom strictly and consistently applied in the practice of management. This paradox can easily be observed in the day-to-day interactions between supervisors and their subordinates and is often associated with the difficulties encountered in the annual performance appraisal (Lefton & Buzzotta, 1977).

The purpose of the present paper is twofold. First, to describe briefly the development of an incentive management program using monetary rewards for Navy civilian employees. This description will focus upon the behavioral principles involved in the development. A more general discussion of the details of the project has been presented in Shumate, Dockstader and Nebeker (1978). The second purpose is to describe the subsequent development of separate programs which used the original program as a model. The development of the later programs was facilitated by the training of the managers and supervisors in the principles of incentive management.

APPROACH

Description of the Measurement System

In its most elemental form, an incentive management system must contain (1) stated goals or objectives, (2) an unambiguous method for determining whether the goals and objectives have been met, and (3) a system for the administration of the rewards so that the rewards are commensurate with the performance. In the system described here these conditions were met by the use of performance standards, performance measurement, and a unique application of the Superior Achievement Award (Federal Personnel Manual, Chapter 451).

Task Studied and Performance Standards. The job studied for the development of the incentive management system was that of the data entry operator, or "key puncher" as it is most generally known. Data entry operators at the locations used in this research use electronic key-to-disc data entry terminals. The information on the disc is later transferred to tape and contains all of the information pertinent to the task being performed, including operator identification and stroke rate in both the write and verify modes.

When the information is transferred from disc to tape it updates a historical file on the particular task being performed. Consequently, historical information for as long as a year is available for each task and for several different operators performing each task. This results in a historical record of the performance of each key entry task across several operators which vary in their performance. For the original test site, hereafter referred to as the pilot site, these historical records were used to establish the standard or expected performance rate for each key entry task. For information on standard development in subsequent applications see Nebeker and Nocella (1979).

Performance Measurement. Fundamental to the successful application of incentive management is the notion of accountability. Such accountability is typically achieved by the development of a thorough performance measurement system. In order to achieve this, management must perform task and work flow analyses of the job being performed and determine the exact role of each human operator in the system. Once the task/role alignments have been established, measures of each task need to be established. Every effort should be made to develop measures which (1) are psychologically meaningful to the human operators and (2) are sensitive to changes in the effort expended on the task. From the management standpoint it is important to be able to relate these kinds of measures to all of the task related activities performed by the person on the job. The extent to which significant portions of the job are not measured can very often determine the success or failure of an incentive management scheme.

The job of key entry at the journeyman level contains five tasks which require accountability. These are: entering data, verifying data previously entered, performing both of these operations on card punch devices, and the preparation of work to be entered. In the program described here, the first two operations were automatically recorded by the key entry terminal. Both stroke count and elapsed time were thus measured by the devices used. Entering and verifying data on cards was measured in a similar fashion, although for some locations only the time spent doing these operations was recorded. Preparatory and set-up time was not strictly measured. Rather, a constant amount of time was allowed during each shift for these activities. This time, plus personal time allowed for breaks, interruptions due to scheduled system maintenance, and the time required for shift-to-shift changes amounted to approximately 1.2 hours. Thus, an operator was expected to be "on line" and in production for 6.8 hours per 8-hour shift.[1] Exceptions to this "standard day" were recorded by the supervisor for each operator. The exceptions allowed were any which were beyond the personal control of the operator, such as system failures, being called away from the terminal to perform other tasks, running out of work, etc.

Determination of Rewards. From the above it can be seen that the work of the data entry operator was measured in terms of stroke rate and time. Specifically, keystroke standards were established in terms of keystroke per hour (KSHR) primarily because this measure of performance rate was one which was readily available and met the previously mentioned criteria of meaningfulness and sensitivity. In addition, production time, the 6.8 hour "standard day", represented an additional production goal that was to be met by each operator. The combination of keystroke rate and production time defines productivity in terms of total strokes (keystrokes/hour X time = total keystrokes)--which is a measure of more meaning to the department supervisors and managers as it is the best measure of the volume of work through the operations section of the department.

Indicies of these three measures were developed in order to facilitate the administration of the rewards. Both keystroke rate (KSHR) and productive time were expressed as percent of their respective standards. Total productivity, i.e., the effects of both speed and time is the multiplicative combination of the indices (Productivity = Keystroke Rate X Productive Time).

The rationale developed for the determination of reward is as follows: When the efforts of the worker exceed production standards, the excess output results in a cost savings to the organization.[2] Under the provisions of most monetary incentive programs, the resultant cost savings are shared with the employee, usually at the rate of 30-70% In the Federal Government, such cost savings are shared at a much lower rate--recommended to be at 10% or less. Determination of cost savings and descriptive examples can be found in Bretton, Dockstader, Nebeker, and Shumate (1978). Savings is ultimately based upon productive performance and the cost per key entry hour. The amount of the reward determined under this program has an upper limit potential of approximately 30% of base salary. In the extreme this means that a key entry operator working at 200% of standard for a year could earn an additional $2850, approximately (cf. Shumate, et al., 1978).

Description of the Management System

The administration of the program required the creation of a new structure within the organizations. This structure, in combination with the measurement system, is referred to as the Performance Contingent Reward System (PCRS)--a title that describes the critical elements of the incentive management domain.

An Incentive Management Coordinator (IMC) was selected from the organization and trained specifically in the administration of the program. This administration was simplified considerably by the development of a management report which utilized the input from the performance recorded by the machines, the standard applied to the jobs, and the bonus as determined from productivity as defined earlier. This report reflected details of performance by task by operator. As an integral part of the system, the IMC was required to present a copy of this report to each operator on the day that the report was produced. Table 1 is an example of this report.

From this report, an operator can determine how his/her stroke rate compares with the rate standards (here identified as TOT EFF), and how the amount of time spent compares with the standard day, here identified as PROD. The PROD EFF column reflects total productivity (RATE X TIME or, in the case presented here, TOT EFF times PROD). Finally, in the last column the dollar amount of the reward, if any. This particular report (not from the pilot site) is run daily and it can be seen that a total of $29.70 in rewards were earned that day--$11.07 by a single person.

Besides the feedback function, the Incentive Management Coordinator was primarily responsible for setting rate standards on new jobs, and processing the payments of the rewards in a timely fashion. The principal points of interface for the latter task were (1) the operator, (2) the Incentive Awards Officer, and (3) the Comptrollers office. An IMC may work only with a first line supervisor and by correspondence with the Incentive Awards Office in order to process the payment. This simple structure, which is supposed to exist for every supervisor working in the Federal Government, is very often hamstrung by several layers of reviewing and approving authority. Every effort was made to keep the PCRS as simple and automatic as possible in order to avoid the debilitating effects of time lag on incentive motivation.

[1] At some locations, control of the work flow was such that only 6.5 hours were required as the "standard day". This most often occurred on the second shift, as operators were frequently required to stop ongoing work to break for a meal.

[2] Implicit in the notion of the standard is that it represents the output that can be expected from an average, fully qualified operator, working at a normal pace.

Table 1

Daily Production Report for Performance and Incentive Earnings

OAL100B KEYPROCESSING OPERATOR ANALYSIS - 790910 MONDAY PAGE 1

OPERATOR	RECORDS		KEYSTROKES				HOURS		KS/HOUR		% EFF		TOT	...PERCENT TIME OF...					PROD	INCENTIVE
	W	V	W	V	T	C	W	V	W	V	W	V	EFF	PROD	IBM	CMC	LV	AD	EFF	EARNINGS
001	1729	516	40337	17113	57450	61	5.04	1.86	7990	9156	110	103	108	97		100			104	.54
002							4.06 AD													
002	639	17	16764	652	17416	3	3.20	.04	5237	15902	89	129	90	97	33	17		50	87	
012	1276	674	34679	18633	53312	77	4 67	2.13	7422	8715	72	75	73	100		100			73	
013							4.33 AD													
013	185	177	8400	6225	14625	21	2.06	.86	4061	7221	60	65	62	90		46		54	55	
019							.16 AD													
019	1345	3470	37765	89173	126938	348	1.91	4.87	19689	18291	239	183	199	102		98		2	202	11.07
020							8.00 SL										100			
021	996	1765	30288	45021	75309	239	2 61	4.22	11564	10658	118	103	109	101		100			110	1.06
024	1876	684	56148	15454	71602	34	5.21	1.31	10776	11796	123	96	117	96		100			112	1.34
025							8.00 AD											100		
026	2188	630	36819	23081	59900	124	3.95	3.32	9309	6935	138	77	110	102		100			112	1.37
030	1176	1050	35642	33600	69242	195	3.28	3.52	10866	9545	103	80	91	100		100			91	
032							8.00 AD											100		
C36	1296	1232	39086	34818	73904	121	3.77	3.20	10356	10863	117	103	111	103		100			114	1.53
039							8.00 SL										100			
040	1457	535	41729	27967	69696	169	4.42	3.05	9438	9160	124	98	113	105		100			118	2.04
045	1263	1976	20041	57607	77848	157	1.63	4.85	12287	11863	159	122	131	95		100			124	2.74
055							8.00 SL										100			
059	2086	1	56162	106	56268	10	6.59	.01	8511	6235	97	47	97	97		100			94	
061	732	1686	25344	45662	71006	147	2.90	4.34	8724	10513	111	127	121	102		100			123	2.54
071	1348	705	42921	19642	62563	152	5.01	1.61	8560	12184	95	132	104	97		100			100	.15
072							4.45 AD													
072	628	598	9712	7647	17359	13	2.67	1.11	3637	6852	96	104	98	112*	40	5		55	109	.48
075							1.00 AL													
075	1165	927	20626	26946	47572	145	2.66	3.80	7736	7083	110	89	97	104		88	12		100	.07
086							8.00 SL										100			
091							8.00 AL										100			
094							4.00 SL													
094	524		22436		22436		3.68		6086		78		78	104		50	50		81	
097	571	1840	20635	62119	82754	244	2.21	4.73	9328	13119	138	152	147	98		100			144	4.77
TOTAL	22480	18683	595534	531456	1127000	2260	67.54	48.90	8816	10866	110	114	111	100	2	64	21	13	111	29.70

General Implementation of the PCRS

The PCRS was tested and evaluated at the pilot site during the year of 1977 (Shumate et al., 1978) and a replication was performed at a second site during the same year, both with similar highly favorable results. As a result of these successes, the Director of the Industrial Activities' Management Information Systems Division of the Naval Sea Systems Command, requested that the Navy Personnel R&D Center prepare the PCRS for export to the remaining six naval shipyards. Time and monetary constraints dictated three major roles for the implementation task: (1) development of the measurement and management systems, (2) management training, and (3) program evaluation. Task 1 was to be undertaken by the field installations, while Task 2 and 3 were the responsibility of NPRDC.

Copies of the production reports and a description of the general management system were supplied to each of the field installations. NPRDC research staff prepared a two-day workshop on incentive management, and developed an evaluation plan for the project. Training for the Incentive Management Coordinators, some supervisors, managers, and Incentive Awards personnel was given during the first week of August, 1978. Approximately one-half of the training time was spent describing the effects of the various parameters of feedback and reward upon work motivation. Of the remaining time, half was spent developing the parallels between theory and management for the data entry job, and the other half was dedicated to problem solving (both with the use of the diagnostic properties of the production report, and with the practical concerns of implementation within the organization).

The Incentive Management Coordinators were then given the task of implementing the system by the beginning of the fiscal year, running the program for 60-90 days in order to establish baselines, and, finally, to introduce the program to the key entry operators at the beginning of the year. Key elements and evaluation criteria for the program were emphasized:

1. Feedback should be timely and provided on an individual basis.

2. Feedback should be informative, i.e., the operators should be aware of the meaning of all of the statistics and indices on the production reports. This would include how the reward was derived.

3. Rate and time standards should be established fairly in order to provide for an incentive effect.

4. The mechanism for processing the monetary reward must be established prior to introduction of the program so that payments can be made without delay (bimonthly payments with a provision for a minimum payment of $25 was recommended).

5. Negative sanctions for low performance should be minimized.

Unfortunately, it was not possible to systematically vary these key elements in order to deal with them in a strict experimental fashion. Feedback with and without a standard was examined in the pilot site (Dockstader, Nebeker, & Shumate, 1977). but time constraints and other practical concerns required that the PCRS be implemented as a complete package for the remainder of the sites. Thus, the evaluation plan allowed only for monitoring of performance rates following implementation, an audit of the standards, and follow-up meetings with the Incentive Management Coordinators to assess the relative effectiveness of the implementation vis a vis the five key elements listed above.

RESULTS

The dependent variables of primary interest in the analysis of the key entry job performance are keystroke rate and productive time. As previously indicated, the measures of these variables are indices, expressed as percentages, which are comparisons of these variables with their respective standards. Thus, the data to be presented will be expressed in terms of percent of standard. The only exceptions to this are data for the pilot site and the second test site. At these locations the performance baselines had been established in terms of keystrokes per hour (KSHR), because of the absence of standards during the baseline period. The results for these two locations are displayed in Figure 1.

Pilot Site Performance

The figures depict pre- and post-implementation keystroke rates and also the indices for productive time and total productivity, i.e., the product of speed and time. Figure 1A portrays the data for the pilot test site. It can be seen from the figure that the average overall keystroke rate for the three months prior to implementation was approximately 8,000 KSHR. Data for a period of one year prior to this time indicates that this baseline is slightly higher than had previously been the case (Shumate et al., 1978). The post-implementation data are averages of the monthly data and reported here in quarter-years. The slight increase during the first quarter has been shown to be cost effective for the organization (Bretton et al., 1978), and the linear increase during the first six quarters represent, at its highest point, a 70% increase in keystroke rate. It can also be seen that productive time increased from 80% to 110% during the first two quarters, then leveled off until the end of the quarter six where it began to increase again as keystroke rate was decreasing. These changes in rate and time reflect turnover within the key entry section--three high performing operators left the section for jobs elsewhere. Their work was picked up (at the cost of time) by other, slower operators. Overall productivity appears to have leveled off during the fifth to tenth quarters at about 140%.

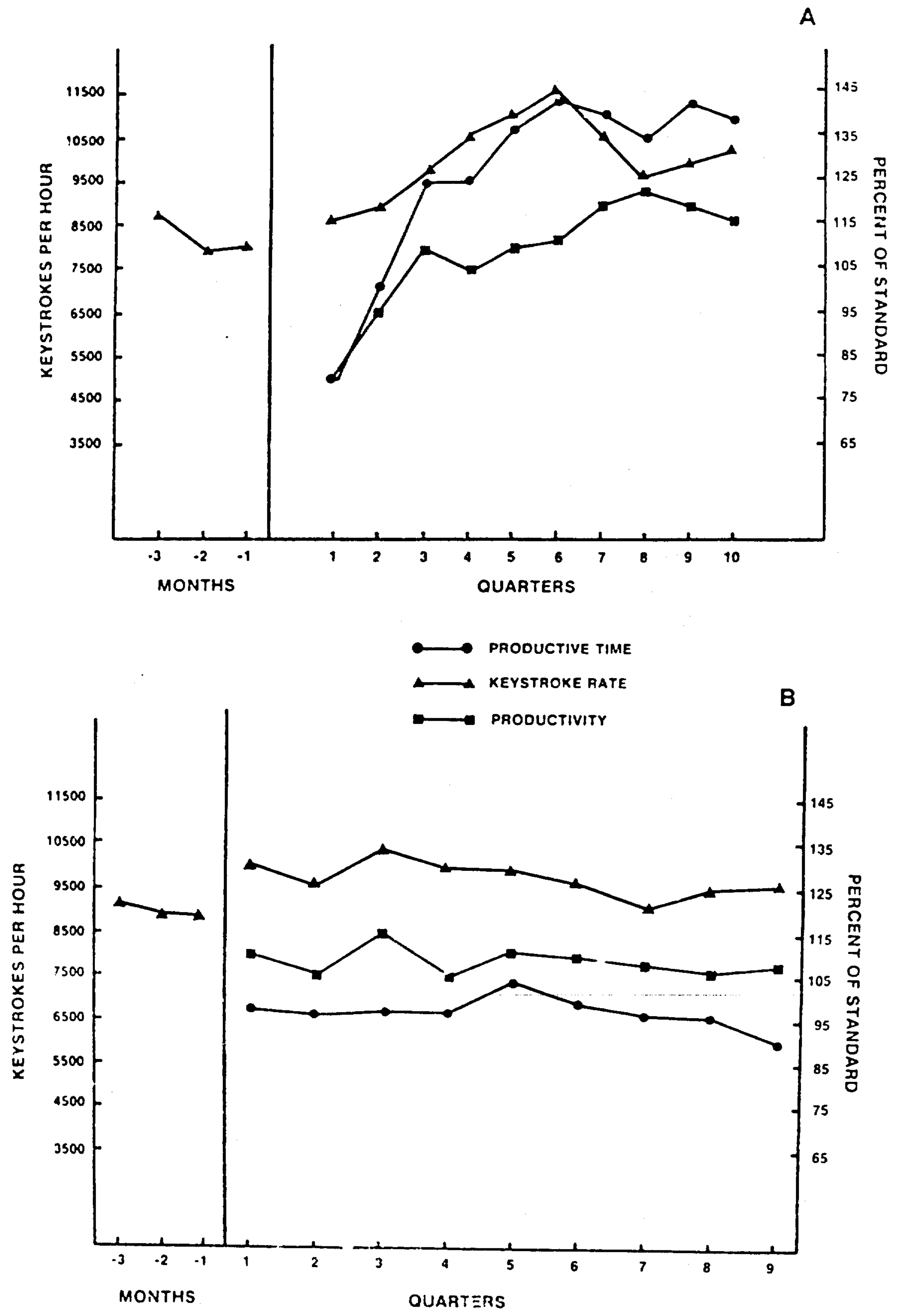


Figure 1. Pre- and post-implementation production statistics for the pilot site (A) and the replication (B). Left ordinate applies only to keystroke rate.

Site B: A Replication

Several months following the implementation at the pilot site, it was decided to test the effectiveness of the PCRS at a second location. The primary purpose was that of a validity test, as a considerable amount of feedback and work flow changes had preceeded the implementation at the pilot site and these effects may have accounted for the changes in performance there (Shumate et al., 1978). In order to logically eliminate such "Hawthorne Effects" we attempted to design the most conservative test for the replication. This was achieved by selecting a second organization which had a well established record of high productivity. Site B was such an organization and had, in fact, an existing incentive management program. The fundamental differences between that existing program and the PCRS had to do with the amount and frequency of the monetary award, and also the productivity standards. At Site B, productivity standards were expressed in terms of keystrokes per day, rather than hourly, and there was a single standard for the day shift and a single standard for the second shift--the difference in standards being a reflection of the different kinds of work performed. The reward was a lump sum of $150 awarded once a year for performance that consistently exceeded standard. This program had been quite successful and had been operating for a considerable time prior to our entrance into the organization.

The supervisor in charge of the existing program was quick to appreciate the differences between his approach and that of the PCRS, but was somewhat skeptical that higher performance rates were possible from his operators. As a progressive manager, however, he was eager to try an experimental program which could have further positive effects on the productivity of his organization. Thus, each job was separately standardized and changes were made in his existing production reports to accommodate the new indices and the determination of rewards. The time standard was not adjusted, however, because the highly efficient work flow system allowed for a 7.1 hour "on line" day.

Figure 1B displays the effects upon keystroke rate. Again, the pre-implementation baseline is stable and representative of performance for the previous year--the overall average being about 8750 KSHR. The changes following introduction of the PCRS was immediate and has been a relatively stable increase of about 1,000 KSHR, an 11% increase in productivity.

Evaluation of Sites A and B

Comprehensive statistical and cost/benefits analyses have been performed on the program developed at Site A (Bretton et al., 1978; Dockstader, Nebeker & Shumate, 1978; Shumate et al., 1978). The need for such an analysis at Site B is questionable because of the stability of the increase and the acceptance of the program by Site B management has not required such a detailed "argument".

A visual examination of the figures does suggest questions of both practical and theoretical interest. For instance, asymptotic performance in terms of KSHR suggests a high similarity between the two locations. However, a comparison of the Productivity measures would indicate that Site A is much more productive than Site B. This is not the case, but is rather an artifact due to the different time standards at the two locations--Site A having a much smaller (and easily attainable) standard day while Site B has the high 7.1 value referred to earlier. Therefore, in the equation Productivity = Rate X Time, there is a greater overall chance that Site A's time figure will exceed 100% than is the case for Site B, thus making the Productivity measure higher.

The theoretical implications of this difference as it relates to work motivation are obvious. First, other things being equal, lower standards allow for higher amounts of monetary reward. Second, lower standards will result in some reward for more persons, i.e., persons of lower ability will be able to earn some reward due to the time they spend working rather than being strictly dependent upon the rate of working. Some managers could view this as a bad feature of this program. However, from the standpoint of incentive management it makes much more sense to be able to appeal to the majority of the work force rather than just the minority of high performers. Motivational theory as well as good insight would predict that unless a goal (i.e., performance level) is perceived to be attainable, the working individual will not strive to reach it. In the example of the key entry task, if the workers cannot achieve the high stroke rates or cannot be in the production mode for 7 hours per day, no reward could be great enough to have lasting effects on their performance.

The multiplicative interaction of Rate and Time in this task can appear somewhat paradoxical to the operator who is capable of high speed. Typically, prior to the advent of this program, operators who were regarded as high performers were those who were capable of high speed. However, it becomes obvious when examining the production reports that some slower operators produce as many or more strokes in a given day, week, or month as the fast operator because of the amount of time they spend "on line". However, this more productive operator may not, in fact, earn any monetary reward when the time standard is as high as 7.1 hours. The effects of this unfortunate inequity is that the slower operator will not receive any reward, will perceive him/herself as incapable of earning the reward, and will thus not be motivated to attempt higher rates of performance. In addition, the high speed operator can be frustrated as well by too high a standard and see that her high rates of performance are not resulting in much payoff. It would be impossible to determine exactly how much influence this frustration could have on the other operators, but it is known to be a factor in the way that the program was perceived at Site A (Dockstader et al., 1978).

The salience of these two standards and their interactive nature becomes more obvious in the sites to be subsequently reviewed. As for the other four important criteria for evaluation referred to in the Approach section--Site A and B received considerable development related to those criteria from the researchers as well as local management. The remaining sites, however, got this information only through the incentive management workshop. The remainder of this paper will address these criteria as they apply to the general implementation of the PCRS at the other sites.

The General Implementation Findings

Only four of the six sites represented at the incentive management workshop have provided data for evaluation. Site visits were made to three of the four locations in order to audit their standards and assess the effective implementation of the five criteria referred to above. Prior to this discussion, two caveats must precede the interpretation of the data: First, no location developed an adequate baseline for the purpose of an unequivocal statistical analysis. The baselines range from less than one month to as many as four--but in each case they were contaminated by prior knowledge on the part of the participant operators. On the surface, this appears to have had little effect on the outcomes and, in any case, it would provide conclusions of a conservative nature--i.e., such contamination usually elevates baselines rather than lowering them.

The second potentially contaminating effect which can be expected to have significant effects on performance is the threat of reduction-in-force (RIF). At the time that most of the sites had implemented the PCRS, headquarters announced that there was going to be a large scale reduction-in-force in key entry. This announcement was passed along to the operators at a later point in time, varying from one location to another. The particular date will be included in the discussion of the data for each site that follows.

Sites C and D. In the four remaining sites, baseline data were available for all three indices. That is, keystroke rate here is an index of keystroke rate compared to the rate standards. These statistics for Sites C and D are presented in Figure 2, which provides a dramatic example of the interactive effects of Rate and Time in the determination of Productivity. In Figure 2C Productive Time increased slightly but remains relatively constant during the eight months following implementation. Keystroke Rate has increased linearly and has been the driving parameter for total Productivity, as can be seen by the close parallel. Figure 2D represents the opposite extreme. In this location keystroke rate has remained relatively constant, while the variations in Productivity have been directly related to changes in Productive Time.

Accounting for the differences between these two sites is also a case of opposites. For the five criteria, there was only agreement on one part of one of them--the rate standards had been correctly set according to the rules developed by Nebeker and Mocella (1979). At Site D regular feedback was not given--it was only provided when an operator had made a reward. Partially as a result of this there wasn't a single operator interviewed who understood the relationship between the Time and Rate standards, nor how the reward was derived from them. At Site C, on the other hand, at least half of the operators and both shift supervisors understood and could verbalize this relationship. The rate standard at Site D was set at 7.0, while it was 6.5 at Site C. It is obvious from production reports there are operators in that organization who are capable of beating the rate standards, but the debilitating effect of the time standard precludes monetary rewards for all but the very fastest. At Site C, payments were promptly made as a part of the bi-monthly paycheck and were paid in full each month. At Site D only two checks had been processed, six full months following introduction of the program, and the checks were for performance 2-3 months past.

One final factor picked up from interviews of operators in Site D had to do with a conflict related to Productive Time. The operators were informed (by memo) that they should attempt to increase their Productive Time; however several reported that they were afraid to run out of work because of fear of losing their jobs by reporting this fact. Thus, any attempts at increasing Productive Time was at the cost of *decreasing* keystroke rate. This obvious bind was not designed into the PCRS, as running out of work was provided as a legitimate reason for *reducing* the work time standard. Running out of work also provides impetus for management to reexamine the work flow system and the role of the supervisor in the distribution of work. It is, after all, *not* the case that there isn't enough work to be done, but rather that the flow of work is interrupted to the extent that it has a negative effect upon Productive Time--unless it is otherwise accounted for.

The announcement of the forthcoming RIF was made at Month 2 for both of these locations. The effect, if any, was transient--although this might account for the large decrease in Productive Time at Month 2 for Site D. In interviews with the operators at all of the locations visited there was an awareness of the impending reduction, but few considered it a serious immediate problem.

Sites E and F. Production statistics for Sites E and F are presented in Figure 3. At the time of this writing, Site E had not been audited so little can be said beyond a description of the statistics presented in Figure 3E. Productivity increases have not been stable, but are, at a a minimum, 15%. Again, as has been the case with most of the sites, the gain is primarily one resulting from increases in keystroke rate rather than in Productive Time.

Productivity at Site F increased following implementation of the PCRS, and the depression at the third and fourth month are interpreted as a reflection of the intention to RIF. Since then there has been a slow but steady increase which appears to have leveled off at 95%--a 20% increase over the baseline figure. The audit conducted at this site revealed that feedback had been regular--daily in this case. The operators generally understood the production statistics and realized that both rate and time were the determinants of productivity and, in turn, rewards. This location, however, had the highest time standard--7.1 hours--and there was considerable frustration expressed by the operators concerning this figure. An examination of the performance rates of individual operators indicated that 50% of them could have been earning some monetary reward if this value was set at the recommended 6.8 hours. As it was, only 25% of the operators were earning some money and only one who was making a significant amount.

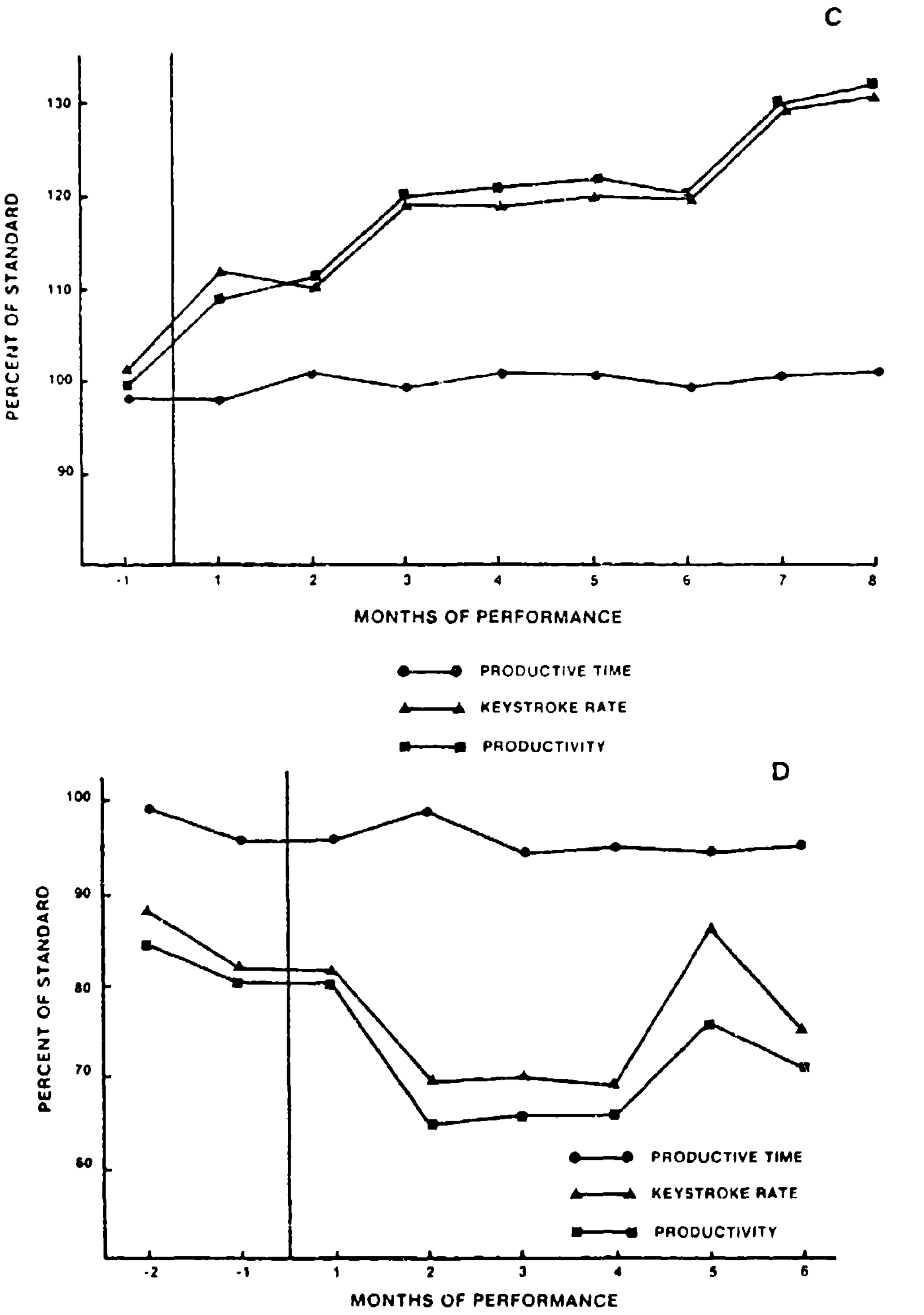


Figure 2. Pre- and post-implementation production statistics for sites C and D, demonstrating the multiplicative relationship between rate and time on productivity.

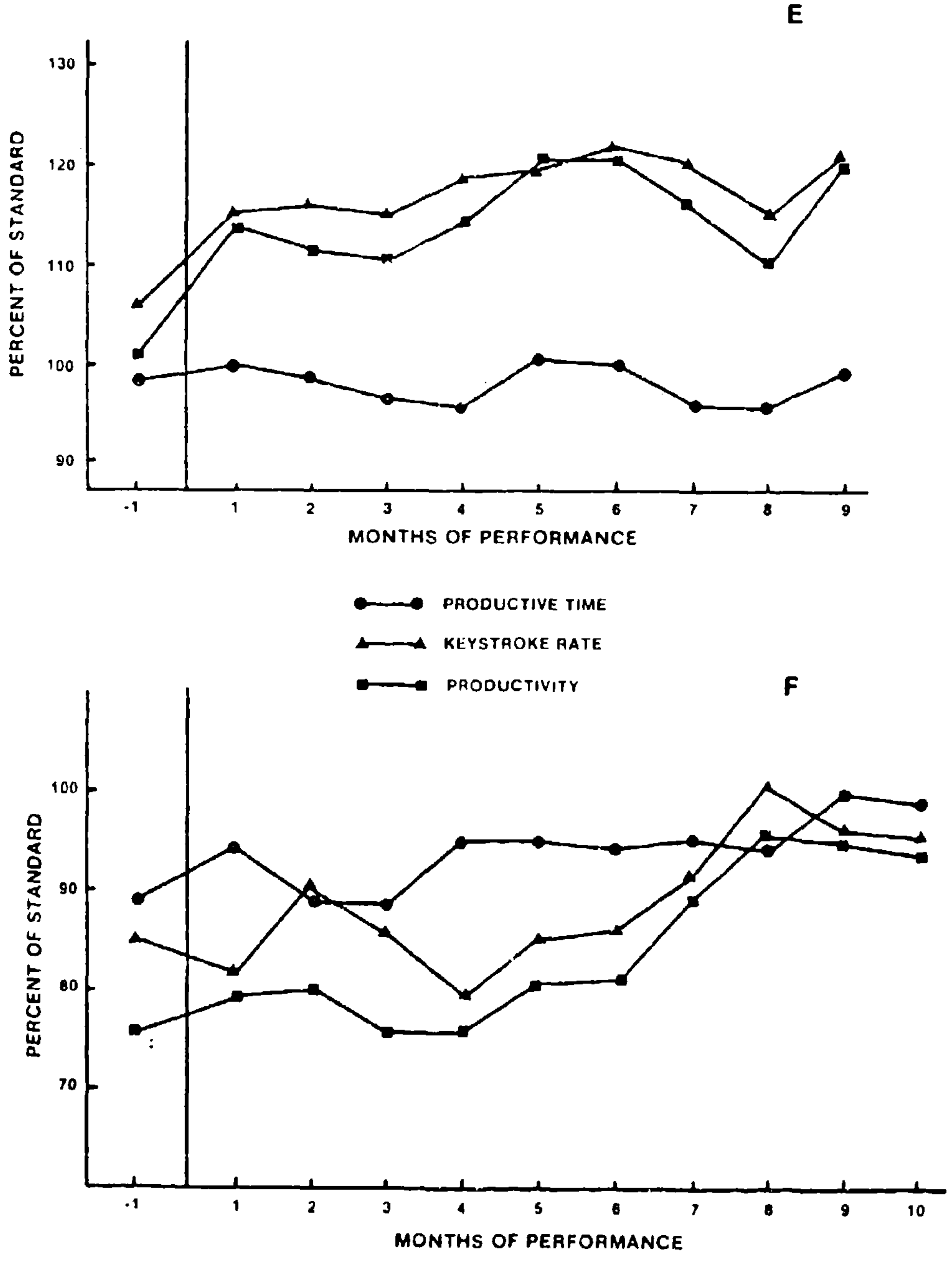


Figure 3. Pre- and post-implementation production statistics for sites E and F.

GENERAL CONCLUSIONS

Six key entry sections were examined regarding the effects of the PCRS on Productivity. In all but one there were significant gains in Productivity which ranged from 11%-55%. In the single location which showed no effects (in fact, net losses) it was found that the time standard was too high for operators to achieve and that several other features of the PCRS were not being practiced, notably the lack of performance feedback and the lack of timely payment of rewards.

The greatest changes in Productivity were found in those organizations that had the lowest time standard, i.e., 6.5 hours. It has been reasoned that this is due to the fact that, other things being equal, this results in (1) greater magnitudes of reward for a given stroke rate and (2) more persons becoming eligible for rewards because it is easier to reach the smaller time requirement. The latter reason is considered the most important because it allows incentives to reach operators who are in the middle of the performance range and will thus include the majority of workers.

It has been observed by the authors during the course of this research program that there has been a general reluctance on the part of management to accept the lower time standard. The reluctance appears to be one which is philosophically based on notions of a work ethic rooted in "a fair day's work" which they equate with eight hours. This idea is firmly associated with compensation, but many managers are apparently unable to divorce it from incentive pay. Paradoxically, some have attempted without success to establish work flow and supervisory practices which would allow for more time "in production" and instead of acknowledging the fact that they cannot raise the value to their goal, they instead set the figure in a way that punishes the worker. The real paradox is that all of the evidence concerning the effects of monetary incentives indicates that the performance resulting from higher bonuses is proportionately greater than that resulting from smaller bonuses. This is to say that the net effect in terms of cost savings is greater with the more generous programs--a finding rather conclusively demonstrated at Sites A and C.

A basic intention of the design of the PCRS was that negative consequences should not be associated with substandard performance, and that high performance would yield rewards. However, implementation of this program does have a punishment contingency. Specifically, the money earned for high performance is a net amount resulting from the summing of both substandard and suprastandard performance. Thus, if a person has more substandard performance during a particular time period than suprastandard, he or she would lose the money earned for the period of high performance. Thus, the withdrawl of a positive reward results in a punishment. In the organizations with high time standards this is exaggerated and is a known and very aggravating fact because of the difficulty of achieving the high time standard, given the constraints of the production system. This problem could partially be alleviated if the production report that presents the net earnings was run more frequently--most "borderline" operators readily acknowledge that it was easier for them to conceive of consistently performing at above standard rates for a day or a week--but a month was too much to make the effort-payoff worthwhile. This is a potentially positive gain in productivity that is stifled to some extent by a punishment consequence.

Ultimately, the cost savings derived from increased individual output come from a reduction in the number of manhours required to perform a fixed amount of work. Thus, unless organizations have an exceedingly large backlog or are in a growth period, savings come from the reduction of personnel. In the key entry task this was easily accommodated because of the large amount of turnover in this job. The pilot site found it easy to absorb the loss of almost one-third of the key entry personnel because of the increased output of the remainder. The implication of this kind of program for other jobs is straightforward: mechanisms currently exist with the Civil Service for continuing monetary rewards for high performance. This fact is very important in the light of anticipated ceiling point reductions and the large number of retirements from Federal Service being predicted during the 1980's. Effective performance-related reward systems like the PCRS are required to meet the difficulties associated with these large changes in the available manpower.

REFERENCES

Bretton, G., Dockstader, S., Nebeker, D., & Shumate, E. Preliminary cost effectiveness analysis, projections, and implementability issues of a performance contingent incentive system. (NPRDC TR 78-13) San Diego: Navy Personnel Research and Development Center, February 1978.

Dockstader, S., Nebeker, D., & Shumate, E. The effects of feedback and an implied standard on work performance. (NPRDC TR 77-45) San Diego: Navy Personnel Research and Development Center, September 1977.

Dockstader, S., Nebeker, D., & Shumate, E. Performance contingent rewards and productivity: A one-year summary of a prototype incentive management system. (NPRDC SR 78-7) San Diego: Navy Personnel R & D Center, April 78.

Lefton, R. & Buzzotta, V. Effective motivation through performance appraisal. New York: Wiley, 1977.

Luthans, F. & Kreitner, R. Organizational behavior modification. Glenview, Ill.: Scott, Foresman & Co., 1975.

Marx, M. Learning: Processes. London: Collier-Macmillan Ltd., 1969.

Miller, L. Behavior management: The new science of managing people at work. New York: Wiley, 1978.

Nebeker, D. & Nocella, J. Keyprocessing performance: A method for determining operator performance standards. (NPRDC SR 79-22) San Diego: Navy Personnel Research and Development Center, June 1979.

Shumate, E., Dockstader, S., & Nebeker, D. Performance contingent reward system: A field study on worker productivity. (NPRDC TR 78-20) San Diego: Navy Personnel Research and Development Center, May 1978.

CONTINGENCY MANAGEMENT: MATCHING NAVY ORGANIZATIONAL PROBLEMS WITH MANAGEMENT CHANGE STRATEGIES

Linda M. Doherty and Robert L. Holzbach

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

Navy managers, both uniformed and civilian, are faced with a variety of operational and organizational problems. In an attempt to resolve these problems and to maximize unit effectiveness, managers may draw upon any one of a number of established management techniques and strategies. Many of these techniques have been effective in the privvte sector in resolving specific managerial problems. However, the application of these techniques directly to Navy settings may not be appropriate given the unique factors encountered in the military, such as the civilian-military personnel mix, the dual authority hierarchy and the personnel rotation system. Also, past research (Broedling, et al., 1977) has indicated that managers often choose the most familiar management technique rather than the most appropriate one for solving a specific problem.

The overall purpose of the work described here is to investigate the appropriateness of existing management techniques in solving specific Navy management problems. The purpose of this paper is to (1) identify and reduce the set of Navy management problems into a meaningful number, and (2) classify management techniques according to their similarity of application. Given this information, Navy managers would be more likely to select the more appropriate technique for a specific problem.

In modern organizational theory the contingency perspective focuses on the relationship among various factors contributing to organizational behavior and outcomes. Early studies indicated that the effectiveness of an organization is dependent on the "goodness of fit" between its structure and the demands of its environment (Lawrence and Lorsch, 1967; Perrow, 1970; Thompson, 1967; Woodward, 1965).

More recently the contingency concept has been developed further in several specific areas. Schmuck and Miles (1971) have represented an organization as a "cube" with 3-dimensions: (1) diagnosed problems, (2) focus of attention (e.g., person, team, etc.), and (3) the intervention (e.g., training, feedback, etc.). According to this model, diagnosis of problems and decisions regarding focus of attention must precede selection of a particular intervention. In other words, the intervention technique should be contingent on the other two factors.

In the area of management, Kast and Rosenzweig (1973) surveyed contingency views of organization and management and attempted to develop a conceptual model incorporating the environment, as well as the overall organizational system. Also included in the model were organizational goals and values, technical, structural, psychosocial and managerial subsystems. The basic contingency application was that a particular organization could be identified on dimensions related to each of these variables and this specific location may indicate the organizational structure to adopt and the corresponding appropriate managerial actions.

Taking the contingency view of management a step further, Luthans and Stewart (1977) propose a three-factor General Contingency Theory of Management. Their primary variables are environment, resource and management. Two variables may interact to produce the following classes of variables: situational (E X R), organizational (M X R) and performance criteria (M X E). The entire system performance is based on the interaction of all three of the primary variables. From the managers' point of view this model suggests that managers must work creatively within the constraints of the environment and resources to improve the effectiveness of the organization.

Using this framework, Luthans and Stewart suggest that a functional relationship of system variables could be constructed, along with a set of diagnostic instruments to pinpoint particular problems in an organization. Problem diagnosis would indicate appropriate interventions to improve performance, depending upon the state of the organization. Thus, management strategies would be selected based on their impact on contingencies of the organization.

These considerations can be used to form the basic framework for testing the effectiveness of management techniques when applied to management problems. The problems and techniques can be organized as the rows and columns of a matrix. Each cell of the matrix could represent a management technique applied to a particular management problem at a particular organizational site. Due to the very large number of potential problems and techniques the subsequent task is to systematically reduce the number of management problems and techniques to be tested.

MANAGEMENT PROBLEMS

The task was to determine relevant Navy managerial problems. Four considerations determined the relevancy of the problems. The first consideration was which level of management (executive, middle, first line) problems should be addressed. The researchers decided that middle managers had the greatest opportunities for influencing operational effectiveness since managers in the Navy are in jobs for two years and can focus on problems that can be solved in that time span. They also have the requisite power, as well as, broader experience than first line supervisors, and are not too distant from working problems.

Second, it was necessary to identify problems the alleviation of which would have the potential for significant impact on effectiveness.

Third, the problems should encompass a wide range, yet occur frequently in organizations.

Fourth, the problems should be sufficiently independent to allow systematic categorization and ranking according to severity and frequency.

Since the early 1960's, management students enrolled at the Navy Postgraduate School in Monterey, California have written about 500 case studies on managerial problems generally found in middle level management in military (Navy), or mixed military/civilian organizations. The two researchers independently content analyzed a sample of these cases (50) to develop a coding scheme reflecting four major aspects of the problem described:: (1) descriptive characteristics (e.g., organizational setting and severity), (2) managerial functions involved, (3) antecedent or precipitating events (i.e., causes), and (4) behavioral manifestations (i.e., results). Then using the variables found to be associated with these aspects of the problem as the framework for coding the case studies, 10 Navy officers and one civilian executive content analyzed the remaining cases and coded the data for computer analyses. After two two-hour training sessions each judge analyzed approximately 40 cases, with about 10% of the cases analyzed by two judges as a reliability check. A total of 454 cases were useful, with 50 discarded because no management problem was evident in the case.

The next step consisted of grouping the problems described in these cases by analyzing their causes and results using two separate quantitative procedures: (1) a principal components factor analysis (Nie, Hull, Jenkins, Steinbrenner and Bent, 1975) with varimax rotation and (2) a hierarchical clustering OSIRIS Computer program (Johnson, 1967). Although both procedures resulted in generally the same clusters of variables, there were still too many clusters for incorporation in our research design. Therefore, the smaller factors and higher level clusters were examined to determine the importance (frequency) of variables associated with them. There was a close correspondence between factors that accounted for a small percentage of the total variance and the variables that occurred least often in a cluster. These clusters were eliminated, leaving the eight clusters in Table 1.

Insert Table 1 about here

The clusters are identified and labeled together with the variables and their frequencies for each cluster. Both causes and results are generally reflected in each of the clusters, except for the cluster entitled "Performance" (#5), where the results are of overriding importance. Although some Navy managerial problems remain static and others change over time, these eight clusters are considered representative of and broad enough to encompass most of these problems. Thus, they were used in the Problem X Technique matrix.

MANAGEMENT TECHNIQUES

The large number of management techniques made it necessary to reduce the number systematically. The reduction was achieved by classification and the method chosen was based on comparisons of the relative similarity of the techniques.

As a part of determining a classification procedure to use, techniques were considered on the basis of (1) their pattern of effects, (2) underlying philosophy, and (3) required implementation activities. It was determined that patterns of effects for techniques cannot be used to establish similarity because such effects have not yet been identified at this stage of the research, nor documented in the literature. Also, since techniques having the same underlying philosophy may differ in terms of both effects and implied management actions, classification on the basis of philosophy would be of litt.e use to managers. Thus, it appears that classifying techniques based on managerial and organizational activities required for implementation would be the most satisfactory for managerial use, assuming that similar actions produce similar organizational results. This approach, however, requires that implementing activities involving a clearly specified sequence of steps or actions can be described well enough to be objective, for instance to be used in various leadership and management training programs or applied by individual managers with appropriate training and minimal outside consultation.

Given this criterion, descriptive summaries for over 50 management techniques drawn from a wide variety of disciplines (e.g., management science and organizational behavior research) were prepared. 30 techniques had clearly defined procedural implementation steps or "critical action requirements." These techniques were then systematically grouped by the following method: Seven behavioral science experts compared all techniques pairwise (N=435) as to the similarity of their critical action requirements. Ratings were made on a 9-point scale. Where 1=most dissimilar and 9=most similar. Finally, to determine clusters of techniques, mean similarity ratings for all pairs were used as input to TORSCA, a nonmetric multi-dimensional scaling program (Young and torgerson, 1967).

The scaling analyses were performed in several dimensions. The 4-dimensional solution provided a close fit when input and output orders of techniques were compared. Six major clusters of techniques were identified by finding distances between techniques in the 4-dimensional space. The identified clusters including labels are presented in Table 2.

Insert Table 2 about here

It is evident that the grouped techniques are consistent with the disciplines from which they were drawn. For example, the organizational development techniques cluster together, while techniques from other disciplines such as operations research are in the management planning group. An exception to this, however, is reward systems and responsibility accounting. They are located in the same cluster, yet do not share the same academic lineage. The common link might be explained in terms of feedback or appraisal systems.

EMPIRICAL MODEL

Incorporating the management techniques with the problem clusters into the problems X techniques matrix results in 8 rows X 6 columns. (See Figure 1.)

Insert Figure 1 about here

Each cell of the matrix represents organizational outcomes when a specific management technique is applied to a particular management problem. The multiple outcomes (O_1. . .O_2) measure changes in behaviors and performance prior to and following the management intervention.

As shown, the rows and columns of the matrix have been prioritized further. Problems had been rated in terms of frequency of occurrence and severity of impact on mission performance. The techniques may be ordered in terms of the likelihood they will be useful in Navy settings. Thus, when considering which cells to test, a segment of the matrix may be identified as high priority, applying the criteria of importance and feasibility.

Although in a simplistic contingency approach a cell of the matrix indicates the usefulness of a management technique applied to a particular management problem, a more realistic approach requires that the entire array of problems may exist in varying degrees in an organization. Furthermore, the impact of a management technique is not necessarily limited to a single aspect of organizational functioning. Broad band assessment of organizational functioning prior to and following an intervention is necessary for evaluation of the effectiveness of a management technique. Since instruments are not yet available to measure all aspects of organizational functioning, the next step would be to develop such an instrument.

Broad-band assessment may be accomplished by assessing each organizational unit on a wide variety of problems. The evaluation of a particular management technique is no longer represented by a cell, but more accurately by an entire column, corresponding to a profile of organizational problems. One classification of problems is performance, and a technique might improve performance problems primarily and secondarily address motivation problems. The predicted and/or obtained effects of each management technique on various problem categories will indicate which techniques have some probability for success.

Therefore, it is important that the assessment of management problems be broad-band to (1) increase the probability of choosing the appropriate technique and (2) evaluate the effectiveness of a technique by measuring all aspects of organizational functioning prior to and following an intervention. Thus, each management site serves to test the hypothesized effects of the selected technique(s) on certain primary problems, and provides further information in order to build a pay-off matrix enabling managers to choose the optimum technique to apply to the solution of their unique array of problems.

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary

A contingency framework of management in which problems of system performance result from an interaction of situational and organizational variables was adapted. This conceptual framework forms the basis of a research model to evaluate management techniques in solving management problems.

Eight clusters of management problems were identified through analysis of case studies drawn from Navy settings.

Six clusters of management techniques were identified drawn from a wide variety of techniques.

These management problems and technique clusters formed the rows and columns of a matrix to serve as the basic research model which permits a number of comparisons to be made. These comparisons are: (a) the effectiveness of a given technique in lessening a set of problems and (b) the technique which is most suitable for solving a particular problem (or profile of problems).

Conclusions

1. A large number of management techniques are available to Navy managers, some are more familiar to them than others, but almost none have been evaluated in a Navy setting.

2. Comprehensive problem diagnosis is critical for technique evaluation.

3. In terms of research and development, the short range outcome is the development of a tested array of management techniques and the long range outcome is a decision model for managers to choose the optimum management technique in solving organizational problems. The former generally compares cells in a given column while the latter requires comparisons of cells across both rows and columns of the matrix. Identifying the optimum management technique is a more difficult problem. It is assumed that to make these comparisons experimental organizations must be matched on important factors requiring long term extensive commitment by numerous organizations.

Recommendations

1. Evaluation studies of management techniques should be conducted in Navy organizations to include a diversity of organizational types, and management problems.

2. The development of organizational assessment instruments needs to be pursued vigorously and integrated into future organizational research efforts to diagnose management problems and evaluate the effectiveness of management techniques.

REFERENCES

Broedling, L. A., Githens, W. & Riedel, J. Development of management techniques inventory (NPRDC TN 77-12). San Diego: Navy Personnel Research and Development Center, April 1977.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

Kast, F. E. & Rosenzweig, J. E. Contingency views of organization and management. Chicago: Science Research Associates, 1973.

Lawrence, P. R. & Lorsch, J. W. Organization and environment: managing differentiation and integration. Boston: Harvard, 1967.

Luthans, F. & Stewart, T. I. A general contingency theory of management. Academy of Management Review, April 1977, 2, 181-195.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. SPSS: statistical package for the social sciences. (2nd edition), New York: McGraw-Hill, 1975.

Perrow, C. Organizational analysis: a sociological viewpoint. Belmont, CA: Wadsworth, 1970.

Schmuck, R., & Miles, M. Organizational development in the schools. Palo Alto, CA: National Press Books, 1971.

Thompson, J. D. Organizations in ction. New York: McGraw-Hill, 1967.

Woodward, J. Industrial organization: theory and practice. London: Oxford University Press, 1965.

Young, F. W., & Torgerson, W. S. TORSCA, a Fortran IV program for Shepard-Kruskal multidimensional scaling analysis. Behavioral Science, 1967, 12, 498.

Table 1

Management Problem Clusters

		Frequency	Percentage
1.	Authority/Responsibility		
	Responsibility/authority not clear	183	40
	Authority not appropriate for responsibility	89	20
	Not aware of change in scope of position	91	20
	Orders received from more than one source	90	20
	Chain of command violated	110	24
	Authority/responsibility conflicts not resolved promptly	120	26
	Required to be both critic and assistant to superior	25	6
2.	Planning and Coordination		
	Inadequate coordination	116	37
	Insufficient coordination	176	39
	Inadequate planning	158	35
	Inadequate administrative procedures	134	30
	Inadequate work policies/procedures	128	28
	Inappropriate individual or work group goals	124	27
3.	Initiative and Motivation		
	Lack of Initiative	53	12
	Inadequate supervision	164	38
	Insufficient motivation	107	24
4.	Conflict, Cooperation, Communication		
	Inadequate communications	268	59
	Public criticism	70	15
	Lack of confidence/trust	128	28
	Poor interpersonal relations	111	24
	Dissatisfaction	154	34
	Low morale	258	57
	Lack of full cooperation/teamwork	119	26
	Insufficient cooperation	109	24
5.	Performance		
	Degraded performance: quantity	77	17
	Degraded performance: quality	180	40
	Degraded performance: efficiency	175	39
	Inadequate work design/flow	60	13
	Degraded performance: timeliness	133	27
	Degraded performance: mission accomplishment	125	28
6.	Supervisory Behavior		
	Supervisor not responsive	90	20
	Discipline problems	58	13
	Degraded performance: inspections	45	10
	Lack of respect for supervisor	116	26
	Poor habitability/appearance of work areas	18	4
7.	Job Skills		
	Insufficient job skills	73	16
	Job-skill match	41	9
	Low retention-high turnover	51	11
8.	External Policies		
	Compliance with external policies	90	20
	Automation	11	2
	Lack of job security	20	4
	Undesirable personnel transfers	42	9

Table 2

Management Technique Clusters

1. Organizational Development
   - Survey feedback
   - Team building
   - Process consultation
   - Grid organizational development
   - Laboratory training
   - Organizational mirror
   - Role analysis
   - Conflict management
2. Decision Analysis
   - Contingency theory
   - Vroom-Yetton decision model
   - Cost-benefit analysis
   - Delphi technique
   - Judgment analysis
3. Management Planning
   - Incremental analysis
   - PERT
   - MIS
   - Operations research
   - Kepner-Tregoe problem solving
4. Industrial Engineering
   - Work sampling
   - Work flow analysis
   - Work simplification
   - Time management
5. Job Redesign
   - Job enlargement
   - Job design
6. Incentive Management
   - Responsibility accounting
   - Management cybernetics
   - MBO
   - Performance appraisal
   - Behavior modification
   - Reward Systems
   - Scanlon Plan

MANAGEMENT PROBLEMS	Organizational Development	Decision Analysis	Management Planning	Industrial Engineering	Job Redesign	Incentive Management
Authority/Responsibility	$O_1 \cdots O_2$	....	....	....	....	$O_1 \cdots O_2$
Planning and Coordination	'					'
Initiative and Motivation	'					'
Conflict, Cooperation, etc.	'					'
Performance	'					'
Supervisory Behavior	'					'
Job Skills	'					'
External Policies	$O_1 \cdots O_2$	....	....	....	....	$O_1 \cdots O_2$

Figure 1. Empirical Management Problems X Techniques Matrix

A RELATIONSHIP BETWEEN PERFORMANCE AND PERCEIVED VALUE OF FEEDBACK

by

Charles v. Durham
Air University
Maxwell AFB, Alabama

Paper presented at the Military Testing Association Annual Meeting, San Diego, California, 16 October 1979

A RELATIONSHIP BETWEEN PERFORMANCE AND PERCEIVED VALUE OF FEEDBACK[1]

Charles V. Durham

Air University
Maxwell Air Force Base, Alabama 36112

INTRODUCTION

The relationship between performance and feedback has long been of interest to specialists working in many areas. Training and tests and measurements experts are interested in this relationship because of the continued research which indicates that feedback improves performance (cf., DeCecco & Crawford, 1974). Teachers are aware of the importance of this relationship because it can be used to not only improve or change performance on specified tasks but can also alter general patterns of social adjustment (McKeachie, 1969). Industrial/organizational psychologists have recently been provided with evidence that the feedback available in work settings is potentially much more valuable and complex than originally anticipated (Greller and Herold, 1975).

In the industrial/organizational context the value of feedback to employees has continually been a voiced and written concern. The promulgated concern however, is supported by a paucity of evidence indicating the true merit of this variable. To quote Greller and Herold,

> performance feedback in the work setting has been described as central to employee training, performance, motivation, and satisfaction. In light of its bearing on so many issues of interest to industrial and organizational psychology, one is surprised at the lack of empirical rigor which has been applied in an effort to better understand it (1975, p. 244).

In organizational settings, recent interest has focused on the need for and importance of feedback in an effort to improve, change, or "develop" the abilities of individuals over long periods of time. This interest has primarily occurred in relation to the need for training and development of persons to fill high level management positions within

[1] Views or opinions expressed or implied are of the author and are not to be construed as carrying official sanction of the Air University (ATC) or the Department of the Air Force.

industry. A technique called the assessment center has been touted as an answer to this need. It is felt by many to allow for formalization and efficient utilization of the performance-feedback cycle in fostering individual growth (Bray, 1977).

The feature of the center process which is considered key for long-term development is the wealth of detailed information available on each candidate subsequent to the center experience (Finkle & Jones, 1970). This information, if provided the individual in a proper manner, could form the substantive base and motivation necessary to change one's behavior to better meet the needs of a given situation or organization. The assessment center literature has little to offer concerning the "proper manner" but other literature is replete with suggestions and evidence, much of which conflicts. Just three of these conflicts will be mentioned.

1. Feedback of specific versus nonspecific (narrative) information.

Most assessment centers provide feedback in the form of narrative summaries (Slivinski and Bourgeois, 1977), yet research indicates that specific and detailed feedback (to include numerical scores) should be provided for maximum effect (cf., Cummings, Schwab and Rosen, 1971).

2. Participation by the recipient of feedback in the feedback process.

Many assessment centers do not plan for verbal or written participation by the receiver in the feedback process (Slivinski and Bourgeois, 1977). Evidence from other sources repeatedly demonstrates, however, that participation on the part of the receiver is of key importance in acceptance of feedback and change resulting from it (Maier, 1973; Miner, 1975).

3. Emphasizing positive versus negative performance.

This problem focuses on choosing what to emphasize during feedback; the positive, the negative, or some combination of the two. Meyer, Kay and French (1965) found that (a) negative feedback (criticism) tended to arouse defensive behavior and (b) positive feedback did not cause change. Gilmore (1974) demonstrated a very negative side of positive feedback. He found that individuals reinforced for positive performance <u>only</u>, tended to lower their goals in order to insure success and thus more positive feedback.

Because of an increasing interest and emphasis on feedback and its affects on performance, this author decided to investigate the relationships between feedback provided via the assessment center method and receiver perceived changes in behavior as a result of that feedback. Although the assessment center process is quite specialized and unique in its application, information concerning feedback available from this research could provide evidence generalizable to other more used situations.

The dependent variables used in this study were

1. Perceived value of feedback received.

2. Perceived changes that occurred as a result of assessment center feedback.

The independent variables were

1. Assessment center attendance only versus center attendance with feedback.

2. Techniques used in the feedback process. These include

   a. the use of narrative summaries versus actual ratings of behavior.

   b. how much the assessee is allowed to participate in the feedback process.

   c. The effects of emphasizing the positive versus the negative aspects of performance.

3. Assessee performance in the center--high or low--as rated by both the assessee and the center.

Independent variable three relates to a consideration which may be of importance but has thus far not been investigated. It is the impact of an individual's performance <u>during</u> assessment on the perceived value of feedback and subsequent performance change. If, for example, an assessee performed extremely well during the center, he/she might feel change was not necessary and would therefore tend to discount or devalue feedback.

The reader should note that the independent variables in this study are "independent" only in the conceptual sense. This research is primarily a correlational analysis and there are no manipulated variables in the experimental sense.

The specific hypotheses tested were:

1. Those individuals who receive feedback from center personnel at the end of the assessment process will score higher on the dependent variables than those individuals who receive no formal feedback but have only self-evaluative information available.

2. An individuals performance during assessment will moderate perceived value of feedback and performance change.

3. Techniques used in presenting feedback should affect its perceived value and importance in causing behavior change.

METHOD

Two forms of a survey questionnaire were developed: one for distribution to assessed individuals employed by various corporations across the country (see Table 1), the second to be completed by personnel responsible for the respective assessment programs.

Insert Table 1 Here

The assessee questionnaire contained 45 questions; 12 questions were asked concerning value of feedback, 4 related to performance change over the short and long term, and 29 were asked for general information and use in follow-on research.

Returned questionnaires were analyzed using correlational, regression and analysis of variance techniques.

RESULTS

As a check on reliability of the questionnaires used in this study, redundant questions were included and correlation coefficients computed within subjects across questions. These coefficients were then averaged to achieve a measure of overall reliability. Question content and reliability coefficients are presented below. All coefficients were significant beyond the .001 level indicating high agreement even though sample size was quite large.

Content	Coefficient
Accuracy of Performance at Center	.44
Numerical Scores Provided at Feedback Session	.75
Were Notes Taken During Feedback Session	.60
Keep a Copy of the Written Report	.77

Hypothesis 1 stated that of two groups of assessees, one which received center feedback and the other which received no center feedback but had self-evaluative information available, the former would both value the formal feedback more and perceive more changes as a result of it.

Of the total sample of 460 people only 19 (4%) said they did not receive feedback. Since sample sizes were so uneven (19 and 441) a stratified subsample (three respondents randomly chosen from each of the seven companies participating in the study) was selected for comparative purposes. Statistical tests were conducted and confirmed that the subsample was representative of the large population.

A one way analysis of variance was performed and Table 2 shows the group means, F-ratios and probabilities of the resulting analysis.

Insert Table 2 Here

Hypothesis 2 stated that those who attend centers, perform well and receive feedback will score higher on the dependent variables than those who attend, receive feedback, but perform poorly. The hypothesis was evaluated by performing a multiple regression analysis using center ratings of performance as predictors of the dependent variables. Results of this analysis were as follows.

Dependent Variable	R	F-Ratio
Value of Information	.46	42.663*
Short-Term Changes	**	**
Long-Term Changes	.22	5.499***

* $p < .001$

** Contribution to dependent variable too small to enter

*** $p < .05$

To insure the results obtained thus far were not artifacts resulting from extraneous variables, a stepwise multiple hierarchial regression was performed using the following as alternative predictors.

1. Company from which respondents came

2. Sex of respondent

3. Months since assessment

The alternative predictors of sex and company had the most effect in this study. Both variables were significant predictors even though the variance accounted for was quite small (R Sq. Change = .01). It appeared, in review of the data, that one of two explanations accounted for this. One company had a large number of female respondents and the ladies may have perceived the feedback as more valuable and useful than males. Or, even though the variance accounted for was small, there may have been a kind of "climate" variable operating. However, when the regression analysis was performed using a worst possible case, e.g., forcing alternative predictors into the regression analysis first, the independent variables still accounted for a significant amount of the variancc. This shows that the obtained effects were not caused by these two variables acting as moderators or mediators on the independent variables.

In regard to Hypothesis III, a stepwise multiple regression was performed using the various techniques for providing feedback as predictors, Table 3 presents the dependent variables, content of the predictor in the stepwise analysis, and the F-ratio for each entry. Significance was determined by testing the multiple correlation coefficients. The last F-ratio for each dependent variable is not significant at the .001 level.

Insert Table 3 Here

DISCUSSION

As for the assessment center as a stimulus for personal development, the majority of assessees agreed that centers did have one positive effect. Over the long-term, attempts were made to improve in weak skill areas, with less effort being expended on those areas considered strengths. This is in opposition to what prior writers felt the results of the center experience would be. Finkle and Jones (1970), for example, espouse emphasizing strengths during feedback because weaknesses are extremely difficult and time-consuming to change.

When the variable "feedback" versus "no feedback" was considered, it was found that the no feedback group reported significantly fewer long-term changes. This tends to show that providing feedback to assessees enhances the possibility of their using that information subsequently. The data provided by a comparison of the groups that received feedback extends this finding. For individuals who attended centers, performed well and received feedback, the perceived long-term changes were greater than for those who attended the centers, performed poorly and received feedback. On short-term changes there were no differences between groups except when self-evaluation of performance was high. In that case there were more short-term changes perceived, regardless of center rating of performance. One interesting sidelight was the fact that the no feedback group recalled their performance during assessment better than those who received feedback. However, the combination of other factors seemed to have an overall negative effect on putting this knowledge to use. These findings tend to show, therefore, for maximum developmental change some form of feedback should be provided regardless of how poorly the person performs.

As for the feedback techniques used in this research, there were several identified as important: a) informal feedback provided by the supervisor; b) a written report given the assessee to review and keep; and c) the atmosphere of the feedback session. The relationship that surfaced can be stated as follows: oral plus written feedback delivered in a positive way is associated with individuals who value the information more highly and make more long-term changes based upon it. This is a key finding in this study, and suggests that feedback in a face-to-face situation enables presentation of material tailored to the individual while the written report allows the retaining of information for future reference and action. The importance of these three factors: a) positive oral presentation of feedback, b) an informal session by the supervisor; and c) a written report cannot be overemphasized in the feedback process.

Another factor which relates to this discussion is the importance of the amount of feedback information provided. In all cases where feedback was provided by both the oral and written means, scores were higher on the dependent variables. This data lends credence to the hypothesis that amount of feedback information is an important consideration in any form of assessment or appraisal (Kim & Hamner, 1976). It is also notable that combining written feedback with formal oral feedback from a center staff member and informal comments from the supervisor, results in providing feedback from 4 of the 5 sources listed by Greller and Herold (1975).

Although the feedback technique of participation did not materialize as important, three comments are appropriate. First, the importance of one participation measure (being asked about the "why" of performance) was mentioned by many assessees as being an indication that feedback information had been tailored to their specific needs. Another participation factor, being allowed to take notes, was shown related to better remembering the information presented during feedback. And finally, only one-third of the total assessees could recall having actively participated in the feedback session, even to the point of simply filling out an end-of-center critique. It appears that most centers do not feel assessee participation is important or necessary.

In summary of the research here presented and as a tentative bridge between this research and other more useful appraisal situations, the following is provided.

1. Participation in a training or testing situation without receiving feedback should be related to

   a. less value placed upon what the experience was to accomplish.

   b. fewer long-term changes in one's life based on the experience.

   c. the same implementation of short-t changes as those who receive feedback.

2. A person who performs well in a situation and receives feedback should, when compared to a person who does not perform well and receives feedback

   a. place more value on the feedback information received.

   b. experience significantly more long-term and short-term changes in performance.

3. The key factors to be considered when presenting feedback to an individual are as follows:

   a. The amount of feedback. This research indicates the more feedback information offered the higher the resulting value placed upon it and the greater the associated change.

   b. The supervisor should be present when feedback is provided. This does not mean the supervisor should present the feedback (C. F., Miner, 1975).

   c. The supervisor should informally discuss the results of the appraisal with the person.

   d. Both oral and written feedback should be provided.

   e. Feedback should be presented in a positive, non-threatening way.

   f. Feedback should adequately emphasize the weaker areas of performance and how they can be improved rather than over-emphasizing the positive aspects or strengths of behavior.

REFERENCES

Bray, D. W. Current trends and future possibilities. In Moses, J. L. & W. C. Byham (Eds.), Applying the Assessment Center Method. New York, Pergamon, 1977.

Belasco, J. A. & Trice, H. M. The Assessment of Change in Training and Therapy. New York: McGraw-Hill, 1969.

Cummings, L. L., Schwab, D. P. & Rosen, M. Performance and Knowledge of Results as Determinants of Goal Setting. Journal of Applied Psychology, 1971, 55, 526-530.

DeCecco, J. P. & Crawford, W. R. The Psychology of Learning and Instruction (2nd Ed). Englewood Cliffs, New Jersey: Prentice-Hall, 1974.

Finkle, R. B. & Jones, W. S. Assessing Corporate Talent: a Key to Managerial Manpower Planning. New York: John Wiley, 1970.

Gilmore, D. C. The effect of knowledge of results and social reinforcement on goal setting performance and satisfaction. Unpublished doctoral dissertation, The Ohio State University, 1974.

Greller, M. W. & Herold, D. M. Sources of feedback: a preliminary investigation. Organizational Behavior and Human Performance, 1975, 13, 244-256.

Kim, J. S. & Hammer, W. C. Effect of performance feedback and goal setting on productivity and satisfaction. Journal of Applied Psychology, 1976, 61, 48-57.

Maier, N. R. F. Psychology in Industrial Organizations (4th Ed.), Atlanta, Georgia: Houghton Mifflin, 1973.

McKeachie, W. J. Teaching Tips: a guidebook for the beginning college teacher. Lexington, Massachusetts: D. C. Heath, 1969.

Meyer, H. H., Kay, E. & French J. R. P. Jr. Split roles in performance appraisal. Harvard Business Review, 1965, 43, 123-129.

Miner, J. B. Management appraisal: a capsule review and current references. In Wexley, K. N. & Yukl, G. A. (Eds.), Organizational Behavior and Human Performance. New York: Oxford University.

Slivinski, L. W. & Bourgeois, R. P. Feedback of assessment center results. In Moses, J. L. & W. C. Byham (Eds.), Applying the assessment center method. New York: Pergamon, 1977.

Table 1

Source of Respondents
Questionnaire Return Rate and
Total Sample Size

Source	Questionnaires Sent	Questionnaires Returned	Return Rate (%)
A Large Multiline Insurance Company	203	150	74
A Small Medical Insurance Company	21	13	62
A Large Manufacturing Company	100	57	57
A Large Oil Company	92	69	75
A Large Auto Manufacturer	125	114	91
A State Operated Assessment Center	33	24	73
A City Operated Assessment Center	75	33	44
TOTALS	582[a]	460	79

[a] Approximately 400 questionnaires were sent to four other companies that stated they would participate in the study but failed to do so.

Table 2

Comparison Between No Feedback and Feedback
Groups on Dependent Variables

Variable	Group	Mean	F-Ratio	P
Value of Information	No Feedback Feedback	- 5.2880 4.2308	26.409	.01
Short-Term Changes	No Feedback Feedback	6.2632 6.2381	0.002	NS
Long-Term Changes	No Feedback Feedback	0.0563 0.0705	5.785	.05

Table 3

Stepwise Multiple Regression Analysis
with Feedback Techniques on Dependent Variables

Dependent Variable	Predictor	R	F-Ratio
Value of Information	1. Informal feedback received from supervisor.	.43	138.40
	2. A summary of performance given in writing.	.46	17.85
	3. How Feedback was given.*	.50	18.83
	4. Written feedback report given assessee.	.51	5.88
Short-Term Change	1. Informal feedback received from supervisor.	.20	18.02
	2. A personnel specialist was present at session.	.22	4.18
Long-Term Change	1. Informal feedback received from supervisor.	.60	236.48
	2. A summary of performance given in writing.	.65	50.11
	3. How feedback was given.	.69	44.69
	4. Written feedback report given assessee.	.71	14.60
	5. Supervisor present during session.	.71	10.46

*Whether feedback was presented in a positive (supportive) or negative (threatening) way.

THE MIXED STANDARD SCALE AS A SURVEY METHODOLOGY

John Meehan, Dorothy Reed, and
Jimmy Thompson

Directorate of Education, The Air University
Maxwell AFB AL 36112

Walter Hines
DCS Education, Air Training Command
Randolph AFB TX 78148

INTRODUCTION

The classic problems facing evaluators and curriculum developers in surveying graduates of educational programs has been rater leniency and the uncertain reliability of individual respondents. The educational programs at the Air University have all faced the ubiquitous problem of graduates who rate with a rather substantial "halo effect." In addition attempts to survey graduates of other educational programs have produced evidence of rather unreliable responses (Meehan & Dowdy, 1977).

The purpose of the present study was to ascertain the applicability of the mixed standard scale as a survey methodology. Originally introduced by Blanz (1965) and further refined by Blanz and Ghiselli (1972) the mixed standard scale was designed as a performance appraisal technique. According to Blanz and Ghisselli (1972) the technique reduced rater leniency and produced satisfactory interrater reliability.

The traditional mixed standards scale consists of triads of statements which are descriptions of traits. Each statement represents a point on an order of merit scale. One statement prescribes a low level of the trait, one an average level, and one a high level. At least one of these triads exists for each measured trait. The following triad was taken from Blanz and Ghisselli (1972).

I. He is a real self-starter. He always takes the initiative and his superior never has to stimulate him.

II. While generally he shows initiative, occasionally his superior has to prod him to get his work done.

III. He has a bit of a tendency to sit around and wait for directions.

The rater is faced with a list of statements, in random order, and is asked to compare each with the performance of the ratee. In each case he indicates if the ratee is better than, equal to, or worse than the statement. The ratee could thus be described as better than the three statements, equal to all three, worse than all three or some combination thereof. In all there are 27 possible permutations each representing a rating for any given trait.

Of these 27 permutations (Table 1) Blanz and Ghisselli (1972) designated 7 as "consistent" and 18 others as "inconsistent." As noted by Saal (1979) two permutations were curiously omitted but have been included in Table 1.

INSERT TABLE 1 ABOUT HERE

According to Blanz and Ghisselli (1972) consistent combinations show no reversals in the order in which the scaled statements are checked. Each of the logical combinations are assigned a numerical value based on an order of merit for that combination. In addition, Blanz and Ghisselli (1972) also assigned numerical values to each of the illogical combinations. More recently Saal (1979) has pointed out some problems with this scoring system. The system does however permit scoring of not only the logical combinations but also the illogical combinations. In addition, by examining the number of logical versus illogical ratings it is theoretically possible to determine the reliability of individual raters.

METHOD

Utilizing an existing survey designed to elicit alumni opinions about the Squadron Officer School (SOS), the authors modified the instrument to the mixed standard format. Initially the survey contained items arranged in a typical Likert format. To the extent possible triads were built around each item in the existing survey. The respondent was asked to respond to each item by indicating one of the following options.

a) The statement is overly critical regarding this aspect of SOS.

b) The statement accurately describes SOS.

c) The statement is overly complementary regarding this aspect of SOS.

The resulting survey contained 40 triads of statements covering areas of the curriculum. Table 2 shows the areas of the curriculum which were covered by the survey. In addition, a limited number of demographic variables were added and the survey was sent to 499 graduates of the school.

INSERT TABLE 2 ABOUT HERE

RESULTS

Usable responses were obtained from 358 graduates and an examination of demographic variables indicated the sample was reasonable representative of the population. Each triad was scored on the 7 point scale suggested by Saal (1979). Mean scores for each triad failed to show obvious inflation of ratings. Table 3 shows the mean value of each triad along with the percentage of consistent ratings. It is obvious that some triads produce relatively more consistent ratings than others.

INSERT TABLE 3 ABOUT HERE

A factor analysis was accomplished on all 40 triads and the results produced 10 distinct factors. These factors are displayed in Table 4. Since only 24 of 40 triads provided significant loadings, at least 16 triads could be eliminated with no loss of information.

INSERT TABLE 4 ABOUT HERE

DISCUSSION

The results of this preliminary effort indicate that the mixed standard scale can be utilized to replace the traditional Likert type format of an attitude survey. The mixed standard scale did in fact isolate areas where respondents provided particularly unreliable ratings. However it was not possible to isolate the cause of these inconsistent ratings utilizing the current data. On the one hand it is possible that the unreliable ratings were the result of rater limitations while on the other hand they may simply be the result of poorly constructed triads.

The relative amount of rater leniency controlled by the mixed standard technique is at this point, uncertain. Air University evaluation staff members were unanimous in their subjective impression that rater leniency was substantially reduced. Only a future study designed to compare the mixed standard scale with a traditional survey, on an item for item basis, will provide a definitive answer.

TABLE 1

	I	II	III	SCORE
		Consistent Responses		
1	+	+	+	7
2	0	+	+	6
3	-	+	+	5
4	-	0	+	4
5	-	-	+	3
6	-	-	0	2
7	-	-	-	1
		Inconsistent Responses		
8	+	+	0	6
9	+	+	-	5
10	+	0	+	6
11	+	0	0	5
12	+	0	-	4
13	+	-	+	5
14	+	-	0	4
15	+	-	-	3
16	0	+	0	5
17	0	+	-	4
18	0	0	+	5
19	0	0	0	4
20	0	0	-	3
21	0	-	+	4
22	0	-	0	3
23	0	-	-	2
24	-	+	0	4
25	-	+	-	3
26	-	0	0	3
27	-	0	-	2

TABLE 2

Curriculum Area	Number of Triads
Attitude	6
Curriculum	4
Speaking Program	5
Writing Program	7
Leadership Program	5
Evaluation	9
Section Commander	3
Miscellaneous Programs	2
Academic Support	4

TABLE 3

Triad	Mean Score	% Consistent
Pre-Post Performance Appraisal	1.40	91
Sports Injuries	1.05	88
Positive Learning Experience	1.42	87
Curriculum Accuracy	1.54	87
Project X	1.13	84
Pre-Attendance Attitude	1.58	84
Video Tape Equipment	1.14	82
Benefits of Speech Program	1.75	80
Criteria For Grading Writing Assignments	1.62	80
Criteria For Grading Speeches	1.63	79
Speaking Critiques	1.33	79
Fairness in Grading Speeches	2.37	78
At-SOS Attitude	1.60	77
Section Commander as Facilitator	1.46	77
Writing Ability	2.53	77
Clarity of Speaking Program Objectives	1.84	77
Writing Critiques	1.42	76
Attendance	1.87	76
Post-Attendance Attitude	1.50	76
Fairness in Grading of Writing Assignments	2.37	75
Program Challenge	1.71	73
Writing Style	2.47	73
Lectures	1.89	73
Critique System	1.36	72
Leadership Evaluation	2.69	71
Clarity of Writing Program Objectives	2.04	65
Correlation of Sports to Leadership	2.54	65
SOS Learning Objectives	1.76	65
Section Commander as Writing Teacher	2.39	60
Wives Program	1.57	59
Quality of Handouts	1.75	57
Distinguished Graduates	2.54	54
Area Books	1.30	53
Curriculum Value	2.22	48
Motivational Impact of Section Commander	2.54	46
Tests and Objectives	2.12	45
Fairness of Tests	2.33	42
Relevance of Sports to Leadership Ability	3.00	42
Emphasis on Grammar	2.37	10
Speech Format	2.46	7

TABLE 4

I ATTITUDE	II COMMUNICATION EVALUATION	III WRITING EVALUATION	IV OBJECTIVE TESTS	V SECTION COMMANDER
Pre-Post Performance Appraisal	Criteria For Grading Speeches	Writing Critiques	Fairness of Tests	Section Commander As Writing Teacher
Post-Attendance Attitude	Clarity of Speaking Program Objectives	Fairness in Grading of Writing Assignments		Section Commander As Facilitator
Positive Learning Experience	Clarity of Writing Program Objectives	Tests and Objectives	Tests and Objectives	Criteria for Grading Writing Assignments
At-SOS Attitude	Criteria For Grading Writing Assignments	Writing Ability		

TABLE 4 cont'd

VI BOOKS/ HANDOUTS	VII LEADERSHIP & SPORTS	VIII SPEAKING	IX SPORTS	X SPEECH FORMAT
Area Books	Correlation of Sports to Leadership	Benefits of Speech Program	Correlation of Sports Leadership	
Quality of Handouts	Relevance of Sports to Leadership Ability	Speaking Critiques	Relevance of Sports to Leadership Ability	Speech Format
Tests and Objectives				

REFERENCES

Blanz, F. Mixed standard scale: A new merit rating method. Unpublished doctoral dissertation, University of Helsinki, Finland, 1965.

Blanz, F., & Ghiselli, E. E. The mixed standard scale: A new rating system. Personnel Psychology, 1972, 25, 185-199.

Meehan, J., Dowdy, R. The Assessment Center As A Diagnostic Tool For Professional Development. Paper presented at the meeting of the American Psychological Association, San Francisco, August 1977.

Saal, F. E. Mixed Standard Rating Scale: A consistent system for numerically coding inconsistent response combinations. Journal of Applied Psychology, 1979, 64, No. 4, 422-428.

THE EFFECTS OF RATER TRAINING AND RANK ON LEADERSHIP ASSESSMENT

Major R.A. Zuliani

The views and opinions expressed in this paper are those of the author, and not necessarily those of the Department of National Defence.

Prepared for: 21st Annual
Military Testing
Association Conference.
Navy Personnel Research
and Development Center,
San Diego, California
15-19 October 1979

Lieutenant-Colonel
G.M. Rampton
Commanding Officer
Canadian Forces Personnel
Applied Research Unit
Toronto, Ontario, Canada
M2N 6B7

ABSTRACT

This study investigated the effects of rater training and rater military rank grouping on the quality of leadership assessment.

The participants were 30 experienced Canadian Forces' Basic Officer Training Course instructors, and 54 inexperienced raters. The inexperienced raters were divided into three training groups. All participants rated six officer candidates who had been recorded on film.

The results indicated that rater training: increased inter-rater reliability; reduced the halo effect, ($p<.03$); increased sensitivity, ($p<.001$); adversely affected relative rater leniency, ($p<.002$); and, did not affect narrative descriptiveness. There were a number of significant differences between officer and NCO assessments: officers had higher inter-rater reliability than non-commissioned officers (NCOs), ($p<.02$); inexperienced officers had less halo error than inexperienced NCOs, ($p<.03$); and experienced officers were more descriptive in their assessment narratives than experienced NCOs, ($p \leq .04$). The results of this study, for the most part, support the conclusions of research literature that rater training and rater characteristics affect assessment quality.

THE EFFECTS OF RATER TRAINING AND RANK ON LEADERSHIP ASSESSMENT[1]

Major R.A. Zuliani

Canadian Forces Personnel Applied Research Unit
4900 Yonge St., Willowdale, Ontario M2N 6B7

INTRODUCTION

All officer candidates in the Canadian Forces (CF) must attend the Basic Officers' Training Course (BOTC). During the 12 week course, leadership ability is assessed primarily by placing the candidate in a number of field situations where the behaviour of the appointed leader is observed and related to 10 critical requirements. These assessments form the basis for evaluating training progress as well as making decisions on the suitability of trainees as officers in the CF. Owing to the importance of these evaluations a continuous effort is made to maintain and, if possible, improve the quality of leadership assessment at BOTC.

Efforts to improve the quality of performance ratings have usually concentrated on: the opportunity to observe relevant ratee behaviour; developing better rating formats; and, rater training (Borman, 1978). The BOTC field exercises are designed to illicit ample leadership behaviour and rating procedures and format have recently been revised. On the other hand, rater training is particularly important at BOTC because the influx of Regular Officer Training Plan (ROTP) university candidates during the summer months means that a large number of temporary staff must be trained yearly. Therefore, of the traditional efforts to improve assessment quality, rater training seemed to have the greatest potential.

Unlike a number of other militaries, in the CF, officer candidates are assessed by non-commissioned officers (NCOs). This study was an opportunity to measure the relative quality of officers' and NCOs' assessments. Thus, the purpose of this study was to estimate the effects of rater training and rank on the quality of leadership assessments at BOTC.

Assessment quality. Borman and Rosse (Note 1) divide criteria used to measure assessment quality, into three classes: accuracy, reliability, and rating behaviour. Accuracy is "the relationship between a set of measurements obtained with a fallible scale of some sort, and a corresponding set of measurements derived from an accepted standard..." (Gordon, 1970, p.367). Accuracy is the most consequential criteria since it is directly related to predictive validity. Assuming that leadership ability is important to an officer's performance, then the more accurately leadership ability is measured, the better the prediction of the candidate's future success as an officer. In the present study, no acceptable standard could be found; thus, accuracy was not used to measure leadership assessment quality.

[1] Much of the information in this paper was previously reported in, The Effects of Rater Training and Military Rank Grouping on the Quality of Leadership Assessment, Report 78-1, Department of Military Leadership and Management, College Militaire Royal, St. Jean, Quebec, J0J 1K0. The research itself was supported in part by College Militaire Royal and Canadian Forces Training System.

The next rating criteria class includes reliability: "a measure of the extent to which a measure remains constant as it is repeated under conditions taken to be constant" (Kaplan, 1964, p.200). Reliability is a necessary but not sufficient condition for accuracy or validity. Unreliability due to assessment error may involve fluctuations in standards by the individual assessors (intra-rater reliability) or differences in standards by different assessors (inter-rater reliability) (Nunnually, 1967, p.208). For most officer candidates, BOTC is the first prolonged contact with the Canadian Forces; therefore, it is imperative that the candidate not feel that the course is passed or failed on the basis of who does the evaluation. That is, there should be a high degree of inter-rater consistency. Inter-rater reliability, because of its relationship to accuracy and its possible impact on candidates, was considered an important measure of assessment quality in this study.

The third class of rating criteria Borman and Rosse (Note 1) term "rating behavior". Examples of this type of criteria are halo: where the rater's overall impression leads to a bias to rate an individual on certain other attributes in a uniformily positive or negative way; leniency: the tendency to rate everyone favourably (Korman, 1971, p.181); and sensitivity: the discriminating power of an evaluator (Kaplan, 1964; p.200); for example, a rater may measure two people as equivalent because the rater lacks the ability to recognize differences in them. Although it is often assumed that an improvement in rating behaviour will increase reliability and accuracy, this does not always occur. Borman (1975) demonstrated that the reduction of halo effect reduced reliability and did not affect accuracy. In the Borman and Rosse study (Note 1), again, halo was reduced by training but reliability and accuracy were not affected. Thus, the relationship of rating behaviour to higher classes of criteria must be demonstrated and not assumed. Rating behaviour was used as a measure of assessment quality in this study.

BOTC assessors must prepare a narrative on the candidate's behaviour during the field exercise. The ability to produce descriptive narratives is thought to improve observation techniques and thus the quality of ratings in a way similar to the diary-keeping which Bernardin and Walter (1977) reported reduced halo and leniency errors.

Rater training. Most empirical research indicates that rater training has a positive effect on assessments. Rater training has been shown to: increase reliability and validity (Bittner, 1967); improve sensitivity (Boyd, Note 2), reduced halo, (Borman, 1975); and, reduce both halo and leniency error (Bernadin and Walter, 1977). In a study by Latham, Wexley, and Pursell (1975) one training strategy reduced halo, contrast, first impression, and similarity effects. (It should be noted that in the Borman (1975) study, the reduction of halo effect was accompanied by a reduction of reliability). It is Bass and Berret's (1972) contention that without training, "errors will creep into the rating" (p.228), regardless of the evaluation method.

The nature of the training can best be discussed in relation to the four-step evaluation process (cf. Borman, 1978):

a. observe the candidate's behaviour and differentiate between behaviour that is relevant and behaviour that is not relevant to leadership evaluation;

b. match each relevant behaviour to a specific critical requirement;

c. grade each matched behaviour; and,

d. weight these grades to arrive at a single score on each critical requirement.

The BOTC evaluation instrument and procedures are meant to guide the assessors through the four evaluation steps. The narrative is a systematic approach to observing behaviour. The definitions of the critical requirements allow raters to select relevant behaviours and match them to the specific critical requirements. In the same way, scale point definitions aid the grading system. Use of individual standards, of course, reduces reliability. Thus, rater training should be directed at systematic observation of behaviour and at the clarification of critical requirement and scale point definitions. Furthermore, this training must be accomplished in a standardized way if evaluations are to be reliable. (Wildman, Erickson, and Kent, 1975).

Other beneficial aspects of training affect all four evaluation steps. Raters should be made aware of the purpose of the evaluations (Bass and Barret, 1972, p.247). Also, the rater training program should make evaluators sensitive to the different rater behaviour errors and to the methods of counteracting these errors (Guildford, 1954, p.280 cited in Bass and Barret, 1972, p.228).

The format of the training program should allow raters to exercise their evaluation skills. Each step of the evaluation process should be practiced and feedback should be given to raters on the quality of their efforts in a workshop format. A final integration of the four steps would complete the training. The workshop format has been found superior to other training formats such as discussion groups (Latham et al, 1975). Thus, the review of the literature suggests that rater training, especially in a workshop format, should increase inter-rating reliability and reduce rating behaviour errors.

<u>Rater characteristics.</u> There are a number of rater qualities which have been shown to affect evaluation. Bayroff, Haggerty, and Rundquist (1954), testing U.S. Army officers, demonstrated that the accuracy and validity of ratings could be distinguished by a rater's final class standing in officers' school. Korman (1971, p.301), inferred that this finding was due to the intelligence of raters. Besides academic performance, performance on the job has also been shown to affect rater's judgement. Supervisors who are better performers tend to be more discriminating in rating their subordinates, while less effective supervisors are generally more lenient raters (Mandell, 1956). According to Taft (1959), raters who share experiences, standards, and outlooks with the ratee are better assessors. He concludes that familiarity with the types of people who are assessed leads to more valid ratings.

A number of studies have shown a relationship between personality characteristics and assessment quality. Gruenfield and Arbuthnot (1969), investigated two independent variables; field independence - dependence (a cognitive style variable) and masculinity, in their relationship to two dependent variables; sensitivity and halo. Field dependent raters, those who were distracted by irrelevant stimuli on the rod and frame test, were "unable to distinguish among the traits and performance of their peers (low sensitivity) or among various characteristics of each individual peer (high halo)" (p.42). "Raters who distinguished sharply among the performance and characteristics of their subordinates are also likely to be field independent and masculine" (p.42). Tagiuri (1961) identified two areas of difficulty related to rater characteristics. One area of difficulty concerns the complex issue of what constitutes a 'good judge' (i.e., the methodological problems associated with determining the accuracy of assessments). The second problem is that rater quality is probably not based on a unitary rater characteristic, but on an interactive combination of a number of rater characteristics. In summary, the literature suggests that, even though the relationship between rater characteristics and assessment is almost certainly not a simple one, rater characteristics do influence assessment quality.

Not many of the rater characteristics, thought to affect assessments, are easily related to military rank grouping. Without measurement, it would simply be a matter of conjecture as to how intelligence, job performance, cognitive style or other personality traits are distributed between officers and NCOs. Certainly, officers should have an advantage over NCOs in terms of their familiarity with the experiences, standards, and outlook of the officer candidate. Other rater variables, such as status and socio-economic grouping, which may be associated with the different rank groupings, have not yet been studied in relationship to assessment although studies of this type have been recommended (e.g., Greenwood and McNamara, 1967).

METHOD

Participants. The participants in this study were 30 BOTC instructors, 17 officers and 13 NCOs, and 54 inexperienced incremental staff raters, 31 officers and 23 NCOs. This study was part of two staff indoctrination courses. One course conducted at CFB Chilliwack and the other at CFB Borden.

Procedure. The inexperienced raters were randomly assigned to three training groups. The control group received a common two hours of instructions on assessment procedures; the regular group received the common instruction on procedures, a two-hour discussion-workshop aimed at the identification and reduction of rater behaviour errors, and a three-hour field exercise where raters had an opportunity to practice and get feedback on evaluating. The extra group received exactly the same training as the regular group except that before the field exercise they received a two-hour workshop which used a videotape recording (VTR) of a BOTC candidate. This allowed a paced session of evaluation practice and feedback of the four steps of evaluation. The regular and control group were shown the same VTR but were given a diversion task that did not involve evaluating the candidate. This was done to ensure that any training effect of the VTR session was due to the rater training that accompanied the VTR, and not due to having seen a preview of the type of presentation used in testing.

All participants rated six officer candidates who had been recorded on 16mm film during an actual BOTC exercise. The film length ranged from 41 to 51 minutes. The [illegible]s used the standard procedure of writing a narrative during the activity and rating the candidate on 10 critical requirements (CRS) using a 5 point scale with 5 being high and 3 a pass. None of the assessors had previous contact with the six candidates. (For the 10 CRS see Appendix A).

ANALYSIS AND RESULTS

According to standard assessment procedures, assessors are to rate only those critical requirements for which there is observable behaviour. This meant that all raters did not rate every candidate on all critical requirements. Missing data occurred mainly on critical requirements 3 (performing effectively under stress), 5 (supporting/cooperating with others), and 6 (seeking/accepting advice). In order to overcome this missing data problem, mean scores were calculated for the individual component (critical requirements 1 to 4), the command component (critical requirements 7 to 10), and the overall score (critical requirements 1 to 10). The two components and the overall score, rather than the individual critical requirements, were analyzed.

Statistical significance tests were not carried out between experienced and inexperienced rater groups because different sampling techniques were used for the two subject populations. Nevertheless, experienced rater data are included in the same table as inexperienced rater data so that "eye ball" comparisons can be made. As far as training level is concerned, the experienced raters could be considered equivalent to the regular trained group with additional "on job training".

Inter-rater reliability. Inter-rater reliability was estimated by calculating an intraclass correlation (Haggard, 1958; Winer, 1971, pp.283-289) for ratings for each subset of raters (e.g., control group officers, regular group officers, etc.,). This ratio represents the reliability of a single rater in the subset. For experienced raters, the inter-rater reliability for experienced officers is higher than for NCOs (see Table 1). For inexperienced raters, the officers inter-rater reliability was higher than for NCOs in 14 of 18 paired comparisons. When comparing the three training groups to each other for each of the three components, results differed for officers and NCOs. Discounting a tie, extra training produced the highest inter-rater reliabilities in 5 of 5 comparisons for officers. With NCOs, the regular trained groups have the highest inter-rater reliability in 6 of the 6 comparisons. Thus, there was a positive training effect but, when compared to regular training, extra training appears to have increased inter-rater reliability for officers and decreased inter-rater reliability for NCOs.

Halo. The degree of halo for a single rater evaluating a particular candidate was calculated by determining the variance across the critical requirements for that candidate's ratings (Borman, 1975). A lower variance indicates a relatively greater halo effect than a higher variance. After the variance was calculated for each candidate for each rater, the mean of the variances for the six candidates for each rater was determined. This mean variance was the rater's halo score.

An analysis of variance of rater military rank grouping for experienced raters and of location x rater training x rater military rank grouping

Table 1

Inter-rater Reliabilities of Individual Component, Command Component and Overall Scores for Subsets of Raters

Training	Officers n	Individual component R	Command component R	Overall R	NCOs n	Individual component R	Command component R	Overall R
Experienced	17	.62	.71	.69	13	.36	.42	.40
Inexperienced				Chilliwack				
Control	5	.51	.63	.57	3	0	.49	0
Regular	5	.60	.72	.67	3	.47	.65	.69
Extra	7	.68	.82	.81	3	.16	.24	.18
				Borden				
Control	3	.35	.77	.71	4	.52	.47	.44
Regular	4	.61	.70	.71	6	.72	.56	.73
Extra	6	.72	.77	.77	4	.41	.52	.53

(2x3x6) for inexperienced raters was performed on the rater halo score.[2] The only significant findings were on the overall halo score. Inexperienced officers had significantly less halo than inexperienced NCOs, (p<.03), and rater training was significant, (p<.03). An orthogonal comparison between means (Winer, 1971, pp.170-177) indicated that regular and extra groups halo scores were significantly higher than those of the control group, (p<.003), suggesting that training reduced halo effect on the overall score.

Sensitivity. Sensitivity refers to the ability to discriminate among the levels of leadership ability of candidates being assessed. Statistically, it is indicated by a significant interaction effect between ratings assigned to candidates and some other independent variable. A repeated measure analysis of variance (Winer, 1971, pp.561-571, 599-603), rater military rank grouping x candidate (2x6) for experienced raters and location x rater training x rater military rank grouping x candidate (2x3x2x6) for inexperienced raters, produced a number of significant interactions. For inexperienced raters, the rater military rank grouping x candidate interaction was significant for the independent component, (p<.03), the command component, (p<.03), and the overall score, (p<.03), (see Figure 1). Figure 1 represents the interaction on the command component. This pattern is similar to the individual and overall score patterns, suggesting that officers discriminate among candidates better (were more sensitive) than NCOs except for very low rated candidates. This rater rank grouping x candidate interaction on the command component was also significant (p<.001) for inexperienced raters. The pattern being almost identical to that of the experienced raters. For the inexperienced raters, there was a significant interaction for training x rater military rank grouping x location x candidate score for both the individual component, (p<.02), and the overall score (p<0.4). These four way interactions indicate that the individual component and overall score are very unstable for the inexperienced raters (i.e., all of the variables seem to have an effect).

2. Critical requirements 3, 5, and 6 were not incorporated in the halo score (overall variance) because there were too many unobserved ratings.

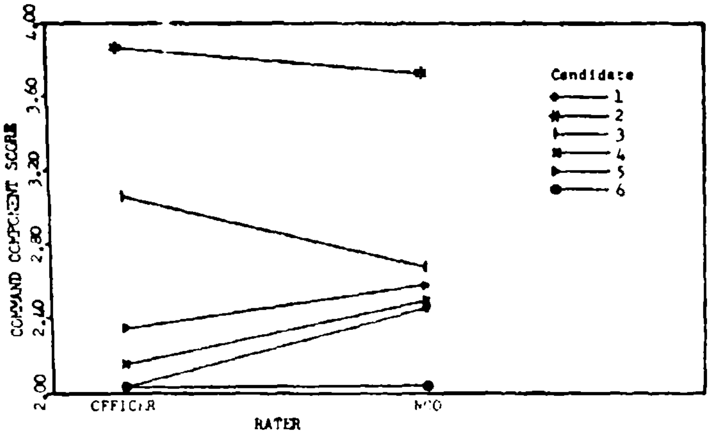


Figure 1: Rater X candidate interaction for experienced raters on the command component score.

Training seems to improve rater sensitivity. The candidate x training group interaction for inexperienced raters was significant ($p < .001$). Trained groups scored better candidates higher and lower candidates lower than the control group. This is especially true for the extra trained group. Trained raters used more of the scale; thus, differentiating more among candidates.

Leniency. Leniency is present when one group of evaluators' ratings are significantly higher than another groups. There was no significant main effect for experienced raters. However, there is evidence that inexperienced raters differed on their ratings at the two locations. The rater training x location interaction, on the command component score was significant ($p<.002$). Virtually no difference exists between the two control groups, Chilliwack $\bar{X}$=2.9 and Borden $\bar{X}$=2.9. However, there are large differences between the trained groups' command component scores at the two locations, Chilliwack regular $\bar{X}$=3.3, extra $\bar{X}$=3.2 while Borden regular $\bar{X}$=2.8 and extra $\bar{X}$=2.6. This difference in candidate ratings for trained groups but not for the control groups suggests that training was a factor influencing Chilliwack personnel to rate candidates more leniently than Borden personnel.

Narrative descriptiveness. A research assistant graded the narratives for descriptiveness according to a four-point scale with 4 being high. The narratives were given to the research assistant in a random order, and he did not know the training group, or rank of the rater. Experienced officers received a significantly higher descriptiveness score, $\bar{X}$=3.3, on their narratives than NCOs, $\bar{X}$=2.9, ($p<.04$). Also, there was a significant military rank grouping x location interaction, for inexperienced ($p<.03$). While the officer narrative descriptiveness score was about the same at the two locations, Chilliwack $\bar{X}$=2.9, Borden $\bar{X}$=2.8 the NCOs at Borden were scored much lower on descriptiveness, $\bar{X}$=2.0, than the NCOs at Chilliwack, $\bar{X}$=2.7. The descriptiveness score for the NCOs at Borden was almost identical to the inexperienced officers' scores.

Observation rates. On a majority of critical requirements, raters scored candidates approximately 100 percent of the time. Notable exceptions to the 100 percent observation rates are critical requirement three (CR3) "performing effectively under stress", $\bar{X}$=64 percent, and the two team component critical requirements, CR5, "supporting/cooperating with others", $\bar{X}$=47 percent, and CR6 "seeking/accepting advice", $\bar{X}$=90 percent. In order to explore the observation rate issue further, the percentage of raters in each training group was calculated. There was a strong tendency for individual raters to either rate all six candidates, or no candidates at all, on CR3 and CR5. The single exception to this pattern is the control group ratings on CR3 where 90 percent of raters evaluated either 5 or 6 candidates. This suggests that the tendency not to make an observation on "performance under stress" is somehow associated with rater training. This pattern did not occur with CR6.

DISCUSSION

Generalizability of findings. A major question that must be asked concerning this study is how well the results derived from a simulated assessment condition can be generalized to the live BOTC assessments. Certainly there are differences between the simulation and the live situation which could have affected the assessment. For example, in the simulation, the raters did not have to show the candidates their assessments or debrief them. Not showing the assessment to the rated individual tends to lower the assessment score (Fournier, Note 3). Fortunately, the conclusions of the study thus far are based on relative measures of assessment quality; e.g., that training increases inter-rater reliability, and that the trained groups in Chilliwack were more lenient than the trained groups in Borden. These conclusions are not based on any absolute measure of assessment quality. Thus, even though it may be true that an absolute measure of assessment quality is not exactly the same for the live and simulated conditions, this does not negate the conclusions.

Nevertheless, it would be desirable to be able to draw some conclusions about inter-rater reliability as an absolute estimate of inter-rater reliability in the live BOTC conditions. Although the degree of bias cannot be measured exactly, research evidence does suggest the direction of the bias. Historical and biographical data on the candidate which is not available in the simulation, tends to affect certain types of raters differently. In the live situation, environmental conditions also tend to decrease inter-rater reliability, because they do not affect every rater the same way. Knowing one's assessments will be monitored tends to increase reliability Reid, 1970; Wildman et al, 1975). On the whole, the simulation assessment inter-rater reliabilities should be higher than the live BOTC assessments. The BOTC raters, especially the officer group compare quite favourably with the six reliability studies reviewed by Borman (1978) in which the intra-class correlations ranged from .36 to .69.

Negative training effects. Both the regular and extra trained groups at Chilliwack rated candidates more leniently than the regular and extra trained groups at Borden, but there was little difference between the two control groups. The likely cause was an aspect of the training that was common to regular and extra trained groups but unique to Chilliwack. This eliminates all but two training sessions: the discussion of assessment errors and the practical assessment exercise. The discussion of assessment errors was given by the same instructor at both locations, who had as a training goal the standardization of training at the two locations. The practical assessment exercise was given by different instructors at the two locations, and is the likely source of the relative leniency effect. The implication is that someone in authority, perhaps a course instructor or a company commander, can dramatically influence an inexperienced assessor's ratings which in turn would have an affect on candidate success rates.

Extra training, which reduced halo and increased sensitivity for both officers and NCOs, had a differential effect on inter-rater reliability for the military rank groups. With the extra training, officers' inter-rater reliability increased, but that of the NCOs decreased. It is possible that NCOs had difficulty interpreting the additional information provided in the video tape training session. At least this interpretation was not uniform across NCOs. This leads one to the conclusion that officers and NCOs differ in their ability to integrate information. Perhaps NCOs are not as well grounded as officers in the leadership model on which the assessment procedures are based. Then NCOs would not be expected to use this officership model as a basis of integrating information in a consistent manner.

Military rank grouping. Even though the sample sizes are only moderate, significant results and consistent patterns point to a fundamental difference in the quality of officer and NCO assessments. There are a number of plausible explanations for this apparent consistent difference:

a. Officers and NCOs may differ on a large number of rater characteristics thought to affect ratings, e.g., shared experiences, standards, outlooks with the candidates; and, cognitive style.

b. NCOs were not given the same opportunity to familiarize themselves with the leadership model and leadership training received by the candidates as did the officers. For example, a four hour segment of the indoctrination course, devoted to leadership training, is for the officers only, and as far as the experienced staff is concerned, only the officers teach leadership in the classroom.

c. It may well be that the differences mentioned in a. and b. operate in combination with each other.

Practical implications. From a practical point of view, this study has shown that evaluation simulation does discriminate among evaluators on a number of assessment quality criteria. Thus, simulation could be used as an evaluator selection device. There is evidence that rater training can increase reliability, decrease halo effect, and increase sensitivity for inexperienced raters. The leniency effect and the observation patterns indicating fundamental disagreements over concept underlying two critical requirements support Wildman et al. (1975) contention that training must be standardized. These negative effects of training and other indications of the inherent instability of subjective evaluations argue for the establishment of a monitoring system as part of any evaluation system. Finally, the failure of the video tape workshop to produce the positive reliability effects in NCOs that was produced in officers suggest that there may not be one optimum rater training strategy. Further research is required to determine the specific rater characteristics which are most important in the rater-training strategy match. The results of this study generally support the conclusions of the research literature: rater training improves assessment quality and rater characteristics, represented in this study by rater military rank grouping, affect assessment quality.

REFERENCE NOTES

1. Borman, W.C. and Rosse, R.L. Format and training effects on rating accuracy and rater errors. Paper presented at the 80th Annual meeting of the American Psychological Association, Toronto,1978.
2. Boyd, J.E. Longitudinal Comparison of Factor Structures in a Police Performance Appraisal System. Unpublished paper, University of Calgary, 1975.
3. Fournier, B.A. Situational Testing as a Selection Instrument: A Review of the Literature (Report 74-4). Toronto, Ontario: Canadian Forces Personnel Applied Research Unit. July, 1974.

REFERENCES

Bass, B.M., & Barret, G.V. Man Work and Organizations - An Introduction to Industrial and Organizational Psychology. Boston: Allyn and Bacon, 1972.

Bayroff, A.G., Haggerty, H.R., & Rundquist, E.A. Validity of ratings as related to rating techniques and conditions. Personnel Psychology, 1954, 7, 93-113.

Bernardin, H.J., & Walter, C.S. The effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Bittner, R. Developing an employee merit rating procedure. In E.A. Fleishman (Eds.). Studies in Personnel and Industrial Psychology, Homewood, Illinois: The Dorsey Press, 1967.

Borman, W.C. Effects of instruction to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 1975, 60, 556-560.

Borman, W.C. Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 1978, 63, 135-144.

Gordon, M.E. The effect of the correctness of the behavior observed on the accuracy of ratings. Organizational Behavior and Human Performance, 1970, 5, 366-377.

Greenwood, J.M., & McNamara, W.J. Inter-rater reliability in situational tests. Journal of Applied Psychology, 1967, 51, 101-106.

Gruenfeld, L., & Arbuthnot, J. Field independence as a conceptual frame-work for prediction of variability in ratings of others. Perceptual and Motor Skills, 1969, 28, 31-44.

Haggard, E.A. Intraclass Correlation and the Analysis of variance. New York: Dryden Press, 1958.

Kaplan, A. The Conduct of Inquiry: Methodology for Behavioral Science. Scranton, Penn: Chandler Publishing Co., 1964.

Korman, A.K. Industrial and Organizational Psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Latham, G.P., Wexley, K., & Pursell, E.D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60, 550-555.

Mandell, M.M. Supervisory characteristics and ratings: A summary of recent research, Personnel, 1956, 32, 435-440.

Nunnally, J.C. Psychometric Theory. New York: McGraw-Hill, 1967.

Reid, J.B. Reliability assessment of observation data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Taft, R. Multiple methods of personality assessment. Psychological Bulletin, 1959, 56, 333-352.

Tagiuri, R. Person Perception. In Lindzey, G. & Aronson, E. (Ed.). Handbook of Social Psychology, (2nd Ed.). Reading Mass.: Addison-Wesley, 1969, 395-449.

Wildman, B.C., Erickson, M.T., & Kent, R.N. Observer agreement and variability of behavior ratings. Child Development, 1975, 46, 520-524.

Winer, B.J. Statistical Principles in Experimental Design (2nd Ed.). New York: McGraw-Hill, 1971.

Appendix A

The 10 Critical Requirements

Individual component

CR 1 Assuming responsibility
CR 2 Displaying initiative/decisiveness
CR 3 Performing effectively under stress
CR 4 Correctly applying knowledge

Team component

CR 5 Supporting/co-operating with others
CR 6 Seeking/accepting advice

Command component

CR 7 Preparation and planning
CR 8 Communicating effectively
CR 9 Directing others
CR 10 Creating high team performance and morale

N.B. The following paper was included in the program of the 21st Annual MTA Meeting, but was not presented at that conference.

OCCUPATIONAL ANALYSIS OF USAF ENLISTED MANAGERIAL, LEADERSHIP, AND COMMUNICATIVE TASKS

Jerry M. Barucky, Captain, USAF

USAF Occupational Measurement Center
Randolph AFB TX 78148

A paper presented at the 21st Annual Military Testing Association Conference, San Diego, California, 17 Oct 79

The views expressed in this paper represent those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense

OCCUPATIONAL ANALYSIS OF USAF ENLISTED MANAGERIAL, LEADERSHIP, AND COMMUNICATIVE TASKS

INTRODUCTION

Each year in an attempt to improve the professional military competence of its personnel, the Air Force enrolls a significant percentage of its force in Professional Military Education (PME) courses. For both enlisted and officer personnel, a series of schools exist which, at various points in a normal career pattern, can be completed either in residence, through homebase seminar programs or via correspondence courses. These common schools are designed to provide a current picture of both the military profession and the aerospace role in national defense, and to build skills and knowledge that will help Air Force people become better leaders and managers.

In an attempt to ensure that the curricula of these programs are pertinent to the needs of those enrolled, the USAF Occupational Measurement Center (OMC) was requested to do a special occupational survey that would identify the leadership, management, and communicative tasks performed by Air Force personnel at each phase of their career. It was hoped that the data from the survey would assist curriculum managers in validating those portions of their curricula.

PROCEDURE

Although the USAFOMC has a great deal of expertise in the analysis of technical jobs, the use of occupational survey methodology to focus on the "soft" skills of managers and supervisors is a relatively new experience. The USAF Human Resources Laboratory had done a similar survey of "officer management activities" in 1964, (Morsh, 1969) and that study served as a guideline for the current effort. Still, many questions existed about the ability to capture meaningful data that would be specific or unambiguous enough to assist curriculum decision makers.

In the overall project we began with the survey of enlisted personnel and used our normal methodology to develop and administer the survey instrument. In the development process this entailed first, fairly extensive research into other job analysis studies used for "professional" or executive-level education. This research was reassuring in that it revealed that many agencies, including the American Pharmaceutical Association, the Canadian Department of National Defense, and most of the US military branches, were currently involved in various similar efforts to identify, objectively, the curriculum needs of their professionals. (Barucky, 1979) Second, detailed interview sessions with groups of experienced personnel were conducted to build a common inventory of leadership, management, and communicative tasks. Third, the inventory was mailed to a fairly large number of personnel (150) at operational bases worldwide so that they could review the statements, critique them for accuracy and clarity, and add any tasks omitted. Finally, validation sessions were conducted with PME school representatives to insure that both the task inventory and the background or demographic questions would provide useable data for curriculum developers. The entire development process took approximately five months and resulted in a list of 264 tasks.[1]

One encouraging aspect of the development process was the surprising degree of consensus, among the enlisted personnel interviewed, about the behaviors to be included in the job inventory. Each interview session began with a lengthy brainstorming session in which the participants listed from 15 to 20 broad areas of responsibility (such as counseling, motivating, or managing resources) that were part of their role as leaders, managers, or supervisors. They then spent the next three days breaking down each of these areas into more specific behavioral statements that could be readily understood by enlisted personnel from all career fields. At the end of each session, the current list generated was compared to the lists generated previously so that omissions, differences, and disagreements in language could be worked out. By the end of the first three sessions it became apparent that 75 to 80 percent of the 200-plus behaviors generated by each group were very similar. Relatively few of the statements proved troublesome or were totally misunderstood by members of other groups.[2]

The survey was administered to a stratified random sample of 11,616 enlisted personnel in paygrades E-1 through E-9. Especially heavy sampling was done in paygrades E-5 and E-9 to allow career-field specific comparisons within those paygrades. Administration took approximately 16 weeks, and resulted in a return of 9,037 useable books. Checks across bases, major commands, and career fields indicated no significant pattern among survey books not returned.

APPLICATION

In using this survey data for curriculum validation, PME curriculum managers from all Air Force major commands gathered at a series of workshops to review the main goals and general objectives for all phases of NCO PME.[3] At these workshops the data was applied in four basic ways:

[1] It is important to acknowledge the fact that some of the statements in the job inventory do not describe discrete, evaluatable, observable behaviors and are, therefore, not really "tasks." However, as these statements were deemed - by the subject matter experts - to be accurate, understandable, or useful from the standpoint of developing curricula, they were left in the inventory. Whatever their title, these tasks/behaviors have provided useful information that helped curriculum developers validate and revise the objectives for enlisted PME.

[2] This agreement among enlisted personnel is somewhat in contrast to the development of the officer job inventory; in the latter effort perceptual differences and extreme concern about the precision of certain statements have led to much more discussion and more interview sessions.

[3] These goals and objectives and the recommended number of hours for each block of PME instruction are listed in an Air Force regulation (AFR 50-29); this document serves as a unifying guideline, which insures that graduates of schools from each major command will have achieved approximately the same skills at each phase of their careers.

(1) to evaluate the proper flow of skill/knowledge development across all five phases of enlisted PME, (2) to help determine, within each block of instruction, the specific objectives that should receive the most emphasis, (3) to identify, within certain phases, specific differences in needs based on career field job requirements, and (4) to determine if the various phases of PME come at the most appropriate times in an NCO's career.

Evaluating the Flow of Skill/Knowledge Development Across all Phases

The questions of when the PME system should first introduce the development of a certain skill or knowledge and the extent (level of learning) to which the skill or knowledge should be mastered were addressed using the percent members performing data. Table 1 illustrates how a comparison of task performance across paygrades assisted in planning the proper sequencing. The workshop representatives considered introducing a skill if at least 30 percent of the students were likely to perform the task employing that skill. Using this rough guideline, it was decided that the skill necessary to write military letters or memoranda should be introduced (to at least a basic level) to senior E-4s or E-5s. It is also evident that a large majority of E-6s and E-7s need to apply this skill, and that, dependent on the amount of instruction provided in the earlier phases of PME, the PME schools attended by E-6s should insure that their students have mastered this skill. On the other hand, Table 1 also indicates that other skills, such as the ability to design or modify organizational structure, need not be addressed until an NCO reaches the grade of E-8. Tasks of this nature would be more appropriate material for a later level of PME.

Determining, Within Each Block Which Objectives to Emphasize

In the second and simplest application of survey data, curriculum developers were able to determine, within various blocks of instruction, some of the specific skills or knowledges to emphasize. For example, from the figures in Table 2, one can see that among the examples of various types of writing skills or formats that one might deal with in teaching written communication to E-6s, it might be more important to insure that the students can write military letters or correspondence than to draft staff summary sheets. The data shows that 54 percent will perform the former task in their jobs, while only 16 percent are likely to perform the latter task.

Identifying Career Field Specific Needs

Another way in which the survey data assisted curriculum developers is through career field specific comparisons made among E-9 personnel. Although E-8 and E-9 personnel are selected to attend the final phase of NCO PME irrespective of career area, a career field comparison indicates that their individual needs for this material based on task performance may be vastly different. Identification of these differences in task performance can allow school personnel to determine both the differences in need and the differences in the experience of students as they enter the school. (For instance, NCOs in career areas such as food service or fuel management are much more likely to need skills in resource management than are NCOs in more technical fields such as band or dental career areas.) With this information, the school personnel can develop elective blocks, modularized or self-paced instruction, or even handouts or reading lists designed to meet these differing needs.

Evaluation of PME Phase Points

The fourth application of the survey data to the overall curriculum validation process was the ability to evaluate the general placement of the five phases of PME within the career span of enlisted personnel. Table 3 illustrates the increasing involvement with leadership, management, and communicative tasks experienced by NCOs as they rise in paygrade. Comparing the differences in involvement at each paygrade with the differences in the amount of material offered in each of the phases of PME corresponding to those grades (Table 4), workshop representatives agreed that the present phasing seems to be logically planned.

SUMMARY

The entire process of validating curriculum objectives according to the actual requirements of the job has been readily accepted and supported by the representatives at each workshop. Although the data has validated a majority of the existing curriculum objectives, numerous revisions have also been suggested. In many cases, the representatives' reaction has been that the revisions recommended actually confirmed their own opinions and some of the opinions expressed on student critiques. They stated, however, that it often takes objective data such as that provided by the survey analysis to convince people that changes are needed.

In summary then, the experience with the enlisted PME project has proved to be quite beneficial. The survey data, while not providing "all the answers," has been helpful to PME managers and curriculum personnel in a variety of ways. In addition, the project itself has shown that an inventory of "soft skill" behaviors can be developed that will discriminate among members of a group based on leadership, management, and communicative task performance. Most important, it has allowed the USAF Occupational Measurement Center to expand its use of occupational survey methodology into a nontraditional area and to open up an avenue for the use of this methodology in an area with a current, crying need - the education and development of professional and executive level personnel.

References

Barucky, J. M. "The Use of Behavior Analysis in the Development of Curriculum for Professional or Executive-Level Education." Unpublished paper, Air Force Occupational Measurement Center, 1979.

Morsh, J. E. "Survey of Air Force Officer Management Activities and Education Requirements." AFHRL-TR-69-38, Lackland AFB TX, Personnel Research Division, December 1969.

TABLE 1

COMPARISON BY PAYGRADE OF PERCENT MEMBERS PERFORMING TWO SAMPLE TASKS

TASKS	E-3	E-4	E-5	E-6	E-7	E-8	E-9
* DESIGN OR MODIFY ORGANIZATIONAL STRUCTURE	8	9	9	14	19	36	41
* DRAFT OFFICIAL LETTERS OR MEMORANDA	15	20	39	54	68	84	89

TABLE 2

E-6's PERFORMANCE OF SAMPLE COMMUNICATIVE SKILLS TASKS

TASKS	PERCENT MEMBERS PERFORMING
* DRAFT OFFICIAL LETTERS OR MEMORANDA	54
* DRAFT INPUTS OR SUPPLEMENTS TO DIRECTIVES	40
* DRAFT MESSAGES FOR ELECTRICAL TRANSMISSION	39
* DRAFT TALKING, BACKGROUND, OR POSITION PAPERS	19
* DRAFT STAFF SUMMARY SHEETS	16

TABLE 3

NUMBER OF LEADERSHIP, MANAGEMENT, OR COMMUNICATIVE TASKS PERFORMED BY 30% AND 50% OF PAYGRADES E-3 THROUGH E-9

PAYGRADE	NO. PERFORMED BY 30% OR MORE	NO. PERFORMED BY 50% OR MORE
E-3	6	4
E-4 (12-48 MOS)	11	4
E-4 (48+ MOS)	20	4
E-5	69	14
E-6	110	41
E-7	161	77
E-8	210	121
E-9	225	139

TABLE 4

HOURS OF INSTRUCTION AT VARIOUS ENLISTED PME PHASES

PHASE	COURSE	HOURS	POPULATION
I	NCO ORIENTATION COURSE	@ 20	E-4 (SR. AMN)
II	USAF SUPERVISOR's COURSE	@ 52	E-4 (SGT)
III	NCO LEADERSHIP SCHOOL	@140	SR. E-4, E-5
IV	NCO ACADEMY	@230	E-6, E-7
V	SR. NCO ACADEMY	@360	E-8, E-9

N.B. The following paper was included in the program of the 21st Annual MTA Meeting, but was not presented at that conference.

VARIANCE WITHIN OCCUPATIONAL FIELDS: JOBS ANALYSIS VERSUS OCCUPATIONAL ANALYSIS

Lt Col J. L. Mitchell and Dr Walter Driskill

Occupational Survey Program
USAF Occupational Measurement Center
Randolph AFB, Texas 78148

INTRODUCTION

Job analysis has been reinvented several times over the last fifty years. As early as 1932, Viteles published his Job Psychograph to identify the personnel requirements of various jobs (Blum and Naylor 1968: 506). In the mid 1930s, the US Training and Employment Service developed a system of worker trait requirements which could be used to relate the requirements of jobs to the characteristics and abilities of workers. This system evolved into the Dictionary of Occupational Titles and a complete job analysis system for use by the Department of Labor (McCormick and Tiffin 1974: 58). In later iterations, the system was modified to be a Functional Job Analysis which classifies all positions in terms of their orientation to data, people, and things (Fine and Wiley 1971). Since the main objective in this system is the classification of jobs, the assumption is made that all the jobs with the same or similar titles have essentially the same content. Thus, any job (or groups of positions) is defined as the mean value on a standard set of descriptors and any variance from incumbent to incumbent is assumed to be idiosyncratic.

In World War II, there was a critical need for a consideration of human variability in the design of weapons systems. This need led, of course, to Human Factors Engineering and an analysis of jobs and positions at a very specific and detailed level. This detailed look at jobs was also called a systems approach, since it viewed the human worker as an integrated part of a man-machine system (Birt and Kimmerling 1978). In this type of system, the range of human abilities had to be accounted for. Seats in aircraft had to be designed for use by people of different heights and weights, or alternatively, the selection procedure had to screen out anyone not within a given range of some variable (i.e., sitting height). This type of concern, however, with multi-factor human variability also generally led to a tremendous wealth of data; so much so that it was overwhelming for personnel decision makers and trainers (Mitchell and McCormick 1979).

A third approach to job analysis was that taken by Flanagan while he was stationed at Lackland AFB, Texas. Colonel Flanagan developed his critical incidents technique (Flanagan 1951) to identify the aspects or factors necessary to success in a job. This listing of the tasks involved in various positions was later refined by Air Force researchers into a comprehensive occupational data analysis system which could deal with all the tasks which a person in a particular occupation might perform. Using modern computer technology, it is now possible to process and summarize extensive amounts of quantitative job information in ways extremely useful to Air Force decision makers, particularly for those involved in the areas of personnel classification, training, and assignments (Morsh 1964, Morsh and Archer 1967, Christal 1974, Driskill 1975). This "task approach" to job analysis today has enjoyed an increasing popularity, not only in the military services but also in the business world where equal employment opportunity concerns make it highly practical (Krzystifiak, Newman, and Anderson 1979).

A fourth approach to job analysis is that of McCormick and his fellow researchers at Purdue University, who have taken a "Worker-oriented Approach" to the study of jobs. Prien and Ronan in their 1971 review of job analysis research findings concluded that McCormick and co-workers had developed a very worthwhile approach to the study of the world of work, one which was particularly useful in terms of quantitative analysis of jobs within and between organizations using a standard questionnaire composed of 189 items (Prien and Ronan 1971). McCormick has, of course, become a world recognized authority in the area of job analysis and the use of quantified job information.

In his now classic review of the area of Job and Task Analysis for the 1976 Handbook of Industrial and Organizational Psychology, McCormick took a very critical look at our state of affairs. He observed that "...the study of human work has generally been more in the domain of the arts than of the sciences. Perhaps to express it differently, the study of the human work (which occupies a major part of man's lifetime) probably has not generally benefitted from the systematic, scientific approaches that have been characteristic of other domains of inquiry... (McCormick 1976:654)."

McCormick's is a rather severe view, and certainly we like to think of the state of the art as being much more advanced than is implied by his review. It is quite possible that his criticism is more a reflection of the lack of basic models and theory building than anything else. While most science is based on the development of good theory (as in physics or even in the area of human behavior, such as social psychology), the whole area of industrial psychology and particularly job analysis has been largely an empirical process. We have no underlying theoretical models which would let us predict human work behavior. Rather, we have generally concentrated on the systematic description of human work, and have invested all our energies in evolving systems which will provide data for practical, real-world decision making.

While the scientist in us may generally agree with McCormick's assessment that the world of job analysis is lacking a good theoretical foundation, as Air Force managers we would assert that science, like everything else, is ultimately judged on the basis of results, not just on the elegance of its theory. And, we can further insist that when judged on the basis of results, the Air Force system of job analysis (sometimes referred to as CODAP - the Comprehensive Occupational Data Analysis Programs), developed by the Air Force Human Resources Laboratory, is in fact one of the most powerful tools available today for gathering and processing quantified occupational information for use in decision making. We believe that the value of the Air Force approach has been recognized in McCormick's review of the area (McCormick 1976), in other academic reviews (of Prien and Ronan 1971), and in the growth of interest in our procedures in both the academic and business worlds (Krzystifiak, et. al. 1979; Cornelius, Hakel, and Sackett 1979, etc.).

The issue of theory versus empiricism is an important one. We would not argue with the need for more basic thinking and writing aimed at building a sound, theoretical foundation under job analysis. Nor would we criticize anyone for building models of job content or job perceptions. These things are needed and certainly we would encourage more work in these areas (Driskill, Keith, and Mitchell 1978).

However, there is a difficulty for the scientist with no prior experience entering the realm of job analysis. Scientists, particularly behavioral scientists, come to this new area with some preconceived ideas about the nature of things. And this can create some severe problems in how they assess the value of various types of information which are typically available in any quantified job analysis system. Today, we would like to discuss just one of these types of information and we hope to show that where an academic approach would discard information as error variance, in the empirical Air Force system we have found that this is one of our most valuable types of information.

VARIANCE IN JOBS VERSUS ERROR VARIANCE

In a typical laboratory experiment, the variance within a group (or treatment) is used, in the usual Analysis of Variance model, as the error term. Thus, within group variance is used to test the significance of mean differences between experimental treatments or groups. Such significance testing is at the heart of most experimental design and is clearly of value in experimental manipulations in the laboratory. This kind of significance testing, however, is also very sensitive to the problems of sampling, the size of the sample, and to the randomness with which subjects are assigned to treatments.

In the real world of work, there is very little practical role for this kind of manipulation of people or for the statistical testing of differences between work groups. Typically, the very large populations of workers in the Air Force make any difference in means statistically significant. The real question is whether there is any practical difference--in terms of the training required for the two groups, the classification structure which best gets the required work done, or the skills and abilities of the individuals required by the different jobs.

To assess the practical significance of any data, one must first define the objective of the analysis. Where the scientist can be satisfied with an objective of just defining if there is a significant difference, the job analyst must be more concerned with the specific objective of a specific study. In this context, the question of statistical significance may have no relevance at all. The question of practical variance in the content of the jobs to be done, however, becomes a matter of paramount importance.

VARIANCE IN AIR FORCE JOBS

If we have two or three thousand Air Force workers who are all repairmen of a particular type of electronic equipment which is used on all types of aircraft we have to decide if they should all be grouped into one occupational classification. Is it the same job on a B-52 Bomber as it is on a KC-135 Tanker aircraft? Does repairing this equipment on an F-4 require the same skills and knowledges as would be required to repair the equipment on the newer F-16 aircraft? Is the job different in overseas locations where no American contractors are available for consultation than in stateside locations where contractors are present or available by phone?

In the Air Force personnel classification system, we have taken the approach that if equipment is generally the same from aircraft to aircraft, if the skills and knowledges are basically the same, and if the time it would take an individual skilled in doing the job on one aircraft to learn to do the job on another system is minimal, then the jobs can be classified as being the same occupation (i.e., Air Force specialty). We end up with more occupational classifications than the Navy has (their Ratings), but fewer specialties than the Army (Military Occupational Specialties or MOS).

In Air Force occupational analysis program, one of our concerns is to study the variation or the range of jobs found within an occupation (specialty). In some studies, we examine the commonality between related specialties specifically to answer the question of whether we could change the structure to reduce the number of occupations (or conversely, we might "shred out" that work to have more specialties). Thus, one of our major objectives is to define the range of variation in jobs--we want to see how many types of jobs there are where there is a practical difference in the tasks being done or in the amount of time spent in performing various sets of tasks.

This concern for the variation in jobs within an occupational field creates a quite different picture for the trainer, for the recruiter, and for the evaluator (specialty knowledge testing). We can no longer take a description of the average job within a specialty and give it to the trainer as a guide for complete, cost effective training. Rather, the trainer must be conscious of all the possible jobs within a specialty that the trainee might be required to do and must prepare the individual to do as many of those jobs as possible. This approach obviously requires a much more sophisticated training system, one which must work closely with the personnel recruiting and assignment systems to insure that the right kind of person is given the right kind of training to do the jobs which need to be done. One might suspect that this is one factor in how our Air Training Command, Air Force Manpower and Personnel Center, and the Air Force Recruiting Service all came to be co-located at Randolph AFB near San Antonio (recently our Office of Civilian Personnel was also established at Randolph as well).

Once we have identified the various types of jobs which are being performed in a given specialty (or set of related specialties), then we can sit down with a group of senior Air Force managers and present the results of our study to them for a decision on what should be done. Let us give you a couple of examples of how this process is working.

AIR FORCE SECURITY POLICE

Recently as part of an occupational survey of the Air Force Security Police career field, we studied the jobs of almost 9,500 law enforcement specialists. If we were to describe the average job of a law enforcement specialist, the most-performed tasks of the composite job description would be as follows:

COMPOSITE JOB DESCRIPTION FOR LAW ENFORCEMENT SPECIALISTS

TASK	PERCENT PERFORMING
Respond to Duress or Alarm Activations	73
Conduct Building Security Checks	72
Operate Vehicle Radio or Public Address System	70
Make Entries on Evidential or Acquired Property Records (AF Form 52)	70
Collect Acquired, Found, or Impounded Property	70
Control or Direct Traffic Other Than in Disaster Areas	69
Conduct Preliminary Investigations of Minor Offenses, Incidents, or Disturbances	68
Make Entries on Statements of Witness (AF Form 69, 70)	65
Perform On-Base Law Enforcement Mobile or Foot Patrols Other Than Missile Security Areas	64
Provide Directions or Information to Visitors	63
Issue Visitor Passes	56

Obviously, the "best-guess" about the tasks that John Doe law enforcement specialist would perform would be the tasks displayed in the composite job description here. The "best-guess" would have to be based on probability. In the case of the job description displayed, the probabilities range from 56 to 73 percent.

Further analysis of the job data for law enforcement specialists reveals the fallacy of the average or "best-guess" job description. Using the clustering procedure which is integral to Air Force CODAP, we found that law enforcement specialists worked in the following seven kinds of jobs:

JOB TYPE	PERCENT	N
General Law Enforcement Patrolman	40	2700
Law Enforcement Operations Technician	25	1650
Installation Entry Controller	10	650
Desk Clerk	5	350
Armory Attendant	5	350
Detention and Corrections Specialist	5	350
Customs Specialist	5	350

A review of the job descriptions for each of these job types reveals that the descriptions for the Law Enforcement Patrolman and Law Enforcement Operations are highly representative of the composite law enforcement job descriptions. As the data show, membership of these two groups accounts for 65 percent of the population of the occupational area.

The job descriptions of the remaining five job types, consisting of 35 percent of the incumbents of the occupational area are not very representative of the composite job description. For example, here are the most-performed tasks in the job description for customs specialists:

CUSTOMS SPECIALIST JOB DESCRIPTION

TASK	PERCENT PERFORMING
Clear Personnel Through Customs	100
Review Customs Declarations	100
Confiscate or Dispose of Agricultural or Edible Materials	100
Report Seizure of Contraband Articles	100
Inspect Military Aircraft for Contraband	93
Search Baggage for Contraband	93
Confiscate Contraband	93
Set Up Customs Inspection Lines for Passengers or Crews Luggage	67

Another example - the job description for the detention and corrections specialist job type:

DETENTION AND CORRECTION FACILITY SPECIALIST JOB DESCRIPTION

TASK	PERCENT PERFORMING
Search Facilities for Unauthorized Articles	98
Inspect Personal Belongings of Personnel in Custody	95
Operate Correction Facility Locks or Doors	92
Admit Personnel Being Placed in Custody	92
Oversee Prisoners, Detainees, Patients, or Visitors	88
Control Entry Into or Movement Within Corrections Facilities	85
Maintain Files of Personnel in Custody	82
Hold or Transfer Prisoners	82

In most of the five less-representative job types, the most-performed tasks in the composite law enforcement description are performed by smaller percentages of incumbents of the job types. In some instances, several of the tasks are not performed at all. Further, the primary tasks of the job types are performed by very small percentages in the composite description.

Implications for using job descriptions which represent the variance of jobs within an occupational field should be fairly obvious--for training, for evaluation, and for career progression. The trainer, faced with only a composite--or "best guess," average--description faces a dilemma. Does he train a little bit on everything--the mile-wide, inch-deep approach? Does he train extensively on everything--the expensive approach? Does he just train the "best-guess" tasks--the most cost-effective approach? Or does he design a training program around the variety of jobs and train personnel for the job which he will perform--the most efficient course of action?

Similarly, testing for promotion and designing career progression programs are difficult without knowledge of the variance of jobs. For testing, unfairness can easily occur, and without the definition of the variant jobs, developing career development paths is virtually impossible.

The discussion thus far has considered training, promotion, and career progression. But what of the other critical aspect of a personnel system--selection? Let's look at an example.

AIR TRAFFIC CONTROL

Several years ago the Air Force Human Resources Laboratory initiated a research project, under the leadership of Dr. Raymond Christal, directed at studying aptitude requirements for selection into Air Force specialties. Data pertaining to the Air Traffic Controller specialty clearly amplify the importance of knowledge of the variant jobs for selection.

An earlier occupational survey identified eleven job types in the Air Traffic Controller specialty. When aptitude requirements for each of these job types were identified, there were large variances, as shown below:

JOB TYPE	ESTIMATED AQE REQUIREMENT[1]
VFR Control Tower	65
Ground Control Approach	70
Precision Approach Radar	72
Mobile Communications	73
Fixed Radar Approach Control	74
Air Traffic Regulation	74
Airport Surveillance Radar	74
Radar Air Route Control	76
Conventional Air Route	78
Special Combat Operations	79
Approach Control Tower	95

Since specific data for a composite job description is not presently available, it is necessary to consider the question hypothetically. Assume that the VFR Control Tower operator is the largest job type, and thus the tasks performed by these personnel would dominate a composite, or average, job description. A poor decision would result if the aptitude requirement for selection were based on this average job. A similar poor decision would result if the Approach Control Tower Operator tasks predominated. In the first case personnel would be required to perform jobs that would stretch their capabilities, while in the second, many would not be sufficiently challenged.

These are just two examples which illustrate the importance of identifying and using data on the variance of jobs. There are many other instances, too, that could be cited to illustrate the importance of identification of variance of jobs--aside from training, testing, and career progression; such data have been essential in studying job difficulty and job satisfaction. In addition, we anticipate it is essential in studying commonality among occupations.

PERSONNEL SYSTEM INNOVATION

As a result of our focus on variation of jobs in an occupation, one of the most significant innovations in personnel classification and training has occurred. Utilization and training conferences involving operators, classifiers, trainers, and occupational analyst for an occupation

[1] Percentages extracted from briefing materials used in a variety of presentations, including one at the Annual Convention of the Classification and Compensation Society, Washington, DC, on 28 July 1978.

convene periodically. These conferences, proceeding from the data base of an occupational survey, arrive at a contract on classification and training. Such an approach has resulted in significant improvements in training and utilization of personnel.

Here's an example--a significant change that would not have occurred without specific information about the variance of jobs in an occupation--in the Loadmaster career field. These are the personnel who supervise the loading of cargo on aircraft and who control the distribution of weight loads in the available space. Total population of this specialty was 2,200 and we surveyed over 1,500 of them or about 70 percent of the Aircraft Loadmasters. When we did a cluster analysis of our responses, we found that the career field was composed of three major clusters and 22 job types.

- Airdrop/Airlift Personnel (N=632)
  - C-130 and C-141 Aircraft
- Airlift Personnel (N=783)
  - C-141 and C-5 Aircraft
- Independent Job Types (N=50)
  - Supervisors
  - Riggers and Packers
  - Flight Examiners
  - Tech School Instructors

If we look at the generalized job description for these personnel, it would appear that about nine percent of their time is spent in preparing for or participating in airdrop operations. However, by looking at the job description for each of the two major clusters, we found that none of the people in cluster II (airlift) were performing airdrop, while almost all of the people in cluster 1 were. Thus, over half of this career field population was not performing airlift, and yet this is an area where training was given in the basic course.

This kind of information was briefed to a utilization and training workshop held at Sheppard AFB TX where representatives of using commands sat down with trainers and others to negotiate the training program. As a result of our briefing on the variance in jobs, the airdrop portion of the basic course was dropped and a unit on airdrop tasks was picked up in the C130 and C141 aircraft familiarization courses conducted by Military Airlift Command. This action saved scarce training dollars and made our training much more effective.

CONCLUSION

Although we are a long way from the first page and the title of this paper, let's go back to it: Jobs Analysis Versus Occupational Analysis. The letter s on the word job is not an editorial error. It's intentional.

The major focus of the Air Force program is on the interpretation of the hierarchical structure of each occupation and in discerning the variance of job content among the various job types within each occupation. By presenting a holistic picture of the job types and their relationships, more efficient personnel management and training programs are possible.

Bibliography

Birt, J. A., and Kemmerling, P. T. Human factors engineering in the Air Force: progress, problems, and diagnosis. Proceedings of the 6th Symposium - Psychology in the Department of Defense, USAF Academy, Department of Behavioral Sciences and Leadership: 69-72, April 1978.

Blum, M. S., and Naylor, J. C. Industrial Psychology. New York: Harper and Row, 1968.

Christal, R. E. The United States Air Force occupational research project. Lackland AFB TX: Occupational Research Division, Air Force Human Resources Laboratory, AFHRL-TR-73-75, June 1974.

Cornelius, Edwin T., III, Hakel, Milton D., and Sackett, Paul R. A Methodological approach to job classification for performance appraisal purposes. Personnel Psychology, 32(2):283-297.

Driskill, W. E. Occupational analysis in the United States Air Force. Paper presented at the Task Inventory Exchange National Symposium on Task Analysis/Inventories, the Ohio State University, Columbus, OH, November 18, 1975.

Driskill, W. E., Keeth, J. B., and Mitchell, J. L. Differential perceptions of Air Force jobs. Proceedings of the 6th Symposium - Psychology in the Department of Defense, USAF Academy, Department of Behavioral Sciences and Leadership, April 1978.

Fine, Sidney A., and Wiley, W. W. An Introduction to Functional Job Analysis. Kalamazoo, MI: The W. E. Upjohn Institute for Employment Research, September 1971.

Flanagan, J. C. Defining the requirements of the executive's job. Personnel, 1951, 28:28-35.

Krzystofiak, Frank, Newman, Jerry M., and Anderson, Gary. A quantified approach to measurement of job content. procedures and payoffs. Personnel Psychology, 1979, 32(2):341-357.

McCormick, Ernest J. Job and task analysis. In M. D. Dunnette (Ed), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally College Publishing Company, 1976, 651-696.

McCormick, E. J., and Tiffin, J. Industrial Psychology. Englewood Cliffs, NJ: Prentice-Hall, 1974 (6th Edition).

Mitchell, J. L., and McCormick, E. J. Development of the PMPQ: a structured job analysis questionnaire for the study of professional and managerial positions. PMPQ Report No. 1, July 1979, Department of Psychological Sciences, Purdue University, West Lafayette, IN.

Morsh, J. E. Job analysis in the United States Air Force. Personnel Psychology, 1964, 17:7-17.

Morsh, J. E., and Archer, W. B. Procedural guide for conducting occupational surveys in the United States Air Force. PRL-TR-67-11. Lackland AFB TX: Personnel Research Laboratory, September 1967.

Prien, E. P., and Ronan, W. W. Job analysis: a review of research findings. Personnel Psychology, 1971, 24:371-396.

N.B. The following paper was included in the program of the 21st Annual MTA Meeting, but was not presented at that conference.

INTERSERVICE JOB ANALYSIS POLICY AND DEVELOPMENT

J. S. Tartell

USAF Occupational Measurement Center
Randolph AFB TX 78148

INTRODUCTION

In August 1972, the Interservice Training Review Organization was established as an informal mechanism within the Department of Defense to review all Service training and education. The purpose was economic in nature - to reduce costs, eliminate duplication - consistent with service readiness. A basic tenet in all of the original agreements lay in the voluntary aspects of the training review efforts.

The Interservice Training Review Organization, as a means for improving training efficiency - read reduce the costs - was driven by a number of factors:

a. The continually escalating costs for all training programs in all services;

b. the paucity of the budget;

c. the requirement for Congressional approval for specific training authorizations; and

d. interest by the Government Accounting Office in the cost and duplication of training among the services.

The efforts of the Interservice Training Review Organization are directed toward increasing the overall training efficiency by which the Services can accomplish the training or education of personnel. However, certain constraints are implied within the sphere of agreement:

a. The interservice program approvals must be based on overall cost effectiveness to the Department of Defense;

b. all interservice programs must be consistent with the readiness, responsibilities, and requirements of the individual Services - read training quality comes before cost effectiveness.

There are also a number of implied benefits attributable to the approach taken by the Interservice Training Review Organization. Active involvement in the training assessment and evaluation system leads to increased communication among the Services. This increase in communication should lead to cross-fertilization of training approaches which will result in improved methods and end-products. The gains in these areas cannot be directly measured in traditional economic terms, but the end result may be more valuable than the tangible sa\ .ngs.

BACKGROUND

The title of this monologue addresses only a single element in the panoply of training in all of the Services. The direction and effort of the Interservice Training Review Organization have, since its inception, been primarily directed toward assessment of curricula. The energies of the working groups established under the aegis of the Interservice Training Review Organization have been aimed at the determination of whether or not the training establishments in the Services have been providing, within relatively the same occupation, equal training. The assumption in this type of analysis was that training course graduates were performing the same jobs or tasks in each of the Services. Such an approach presents immediate problems in terms of vocabulary, classification structure, personnel utilization, and a number of other areas. Despite the problems arising in the areas listed, a number of substantial training revisions were accomplished.

ORGANIZATION

Before presenting a detailed picture of the interservice job analysis approach, an overview of the structure within which the functions were accomplished is deemed appropriate. The Interservice Training Review Organization has undergone a number of structural revisions since its inception. The latest structure, proposed in June 1979, establishes an Executive Board to outline the policy of interservice cooperation in the training arena. The Executive Board is composed of the Service chiefs of training.

Reporting directly to the Executive Board, and implementing the established policy, is the Steering Committee. The members of the Steering Committee also serve as chairmen of the five committees responsible for implementing various aspects of interservice cooperation.

The five committees are:

a. Training and Education Committee

b. Training Management Analysis Committee

c. Training Development Committee

d. Plans and Resource Requirements Committee

e. Health Care Training Committee

Each of these committees has a number of subcommittees responsible for specific aspects of the training or education process.

For the purposes of this discussion, the focus will be narrowed to a single subcommittee operating within the purview of the Training Management Analysis Committee - formerly called the Curriculum Committee. The title of this subcommittee, Interservice Occupational Task Analysis Program Subcommittee, immediately implies the purpose for its existence. The Subcommittee originally was formed under the purview of the Instructional Systems Development Committee.

The Interservice Occupational Task Analysis Program Subcommittee was formed in 1973. The original membership included personnel responsible for collecting and analyzing occupational survey data in each of the Services and each had the authority to commit resources for interservice projects. The primary purpose of the Subcommittee was to foster interservice communication in the area of occupational or job analysis. The basis for the interservice communication - and early cooperation which predates the establishing of the Subcommittee - lay in utilization by all of the Services of the Comprehensive Occupational Data Analysis Programs (CODAP). This common base in programming allowed for discussions to occur in terms common to all representatives despite Service differences.

INTERSERVICE OCCUPATIONAL TASK ANALYSIS PROGRAM

SUBCOMMITTEE: POLICY ESTABLISHMENT

From its inception, the members of the Interservice Occupational Task Analysis Program Subcommittee agreed that interservice cooperation was a desirable and accomplishable objective. The implementation of this philosophy has taken a number of years and several false starts.

The earliest meetings of the Subcommittee established a number of philosophical guidelines which have provided impetus for continued exploration in the interservice arena. These guidelines are listed below:

a. Joint service task development and analysis is feasible;

b. joint task development will be more costly than individual service development, however, there exists greater potential than for single service analysis;

c. results of joint analysis can be applied to both training development and training evaluation; and

d. considerable benefits to each service's analysis program accrue from cooperative efforts.

In addition to these guidelines, the Subcommittee recognized the need to establish some methodology for accomplishing interservice occupational surveys, and a format for getting results applied. These early discussions led to the first interservice occupational survey projects.

In the 1974 and 1975 time frame, three interservice projects were initiated. The first of these was an occupational survey of Explosive Ordnance Disposal personnel. For these specialties there existed a single interservice school and school personnel were intimately involved in the project. The task list, though printed in separate books and administered separately by each Service, was identical. The final analysis was delivered to the interservice school and was utilized in the modification of the course of instruction.

The second project was an occupational survey of Cooks and Bakers in the four Services. For the Cooks and Bakers project, there was no common interservice training, thus, members of the training establishment were not totally involved in the development of the task list. Despite the seeming commonality on the surface, the development of the task list was a long drawn out, often confusing, process. Despite original agreements, the final task list, as printed by each service, was not identical. In addition, incompatibility of computer input resulted in failure of the project.

The third project was an after-the-fact analysis of occupational survey data for the Air Traffic Control specialties. Each Service had constructed their own task list and had gathered data. A panel consisting of occupational analysts and subject matter experts from each Service was convened. The resulting analysis yielded some areas of commonality among the Services, but the problems raised by different task lists written at different levels of task specificity caused the findings to be of marginal applicability.

The consensus reached after the three projects was that interservice occupational projects were feasible. However, the process would not be simple to establish and would require strong management intervention. With that consensus, the interservice occupational survey was placed on the back burner for approximately one year. During this period, much of 1976 and 1977, the Interservice Training Review Organization retreated to a position of assessing - evaluating - the accomplishments to date and establishing policy for continued interservice cooperation.

During the summer of 1977, the Interservice Training Review Organization directed the Interservice Occupational Task Analysis Program Subcommittee to explore methods of interservice job analysis through common occupational surveys.

On 21 July 1977, the Interservice Occupational Task Analysis Program Subcommittee met in San Antonio, Texas. What made the gathering unusual was the presence, by invitation, of representatives of the training establishments of each Service. The purpose for the meeting was, first, to develop a joint position on the need for interservice job analysis, and second, to establish procedures for accommodating the requirement for interservice occupational analysis.

The representatives of the training establishments concluded that interservicc occupational survey data were needed to aid in interservice training decision-making processes. These interservice occupational survey projects should be accomplished using common job inventories. Priorities and time frames were to be recommended by the appropriate Interservice Training Review Organization committees, either the Curriculum Committee or the Medical-Dental Training Committee, and reviewed by the commands responsible for training.

Following the agreement of the training representatives that interservice cccupational surveys are necessary, the members of the Interservice Occupational Task Analysis Program subcommittee established the following guidelines for accomplishing interservice occupational surveys:

a. All Services will support development of common job inventories and data collection to support interservice training decisions;

b. priorities established by the Interservice Training Review Organization must allow sufficient lead time for individual Services to adjust their ongoing survey program schedules;

c. any single interservice occupational survey project may not necessarily involve all Services;

d. every interservice occupational survey project requires a common job inventory - an identical list of tasks - and a common completion date for the availability of data; and

e. funding for interservice projects will remain an individual Service responsibility.

The guidelines and a list of occupational fields for which interservice analysis appeared appropriate (prepared by the Curriculum Committee) were forwarded for consideration by the Steering Committee and Executive Committee of the Interservice Training Review Organization.

The Interservice Training Review Organization in Executive Committee Order Number Ten directed the Interservice Occupational Task Analysis Program Subcommittee to examine the total computer field (excluding tactical or weapon system computers). Thus began the present effort - an occupational survey of enlisted computer operators, programmers, and analysts from the four Services.

THE COMPUTER PROJECT: INVITATION AND DEVELOPMENT

In January of 1979, the Interservice Occupational Task Analysis Program Subcommittee met to review the guidelines for interservice occupational surveys established in 1977, and to determine policy for completion of the occupational survey of computer operators, programmers, and analysts. The following points were agreed to by each Service representative:

a. The occupational specialties for the project were specifically defined: USA-MOS74D, 74F, 74Z; USN-Data Processing (DP) rating; USMC-Occupational Field (OF) 4000; USAF-AFSCs 511X0, 511X1;

b. use of a common task list, identical in number and order of tasks and a number of common background questions;

c. provision of inventory development resources to prepare an interservice job inventory and analysis resources to prepare the final report;

d. designation of the USAF Occupational Measurement Center to accomplish the analysis of the interservice data;

e. establishment of project milestones:

(1) initial task development meeting - March 1979

(2) final task development meeting - April 1979

(3) Service initiation of data collection - June 1979

(4) data provided to USAF Occupational Measurement Center - October 1979

(5) analysis of data and preparation of report - December 1979

(6) brief Interservice Training Review Organization - January 1980

f. the briefing for the Interservice Training Review Organization should address at the minimum the following information:

(1) similarity of jobs and tasks among services;

(2) differences among services based on uniqueness of mission, career, progression patterns, and experience level of personnel;

(3) a judgment, based on occupational criteria, whether the standardization of occupations is supportable.

The representatives from each Service further agreed that each should begin development of a task list. For the Air Force this proved to be an item of little consequence. Occupational surveys of computer operators and programmers had been accomplished thrice, the last time in 1977. The task list had been updated after each administration and was considered relatively current, requiring only validation. For the Army the process was somewhat more involved because they had not yet done an occupational survey in these specialties. However, the construction of a task list had been initiated in response to a request from the Army's training establishment. Neither the Navy nor the Marine Corps was as fortunate. Both had to start the project from scratch. To further facilitate the task list development efforts, each Service agreed to exchange task lists prior to the first consolidation meeting - this later proved to be an extremely valuable exchange.

The first task consolidation meeting was held in March 1979 in Washington D.C. That location was chosen because it involved significantly less money to meet there than at any other location. The primary purpose of the meeting was to consolidate into a single list the task lists of each Service. To facilitate this consolidation the Marine Corps representatives had keypunched each Services input and had prepared a single master listing. This single listing contained more than 2,000 entries. The consolidation effort required individual evaluation of each task statement. To accomplish this review of the 2,000 tasks, each Service was represented by an inventory development specialist and at least one subject matter expert. The project officer appointed by the Interservice Occupational Task Analysis Program Subcommittee chaired the meeting and resolved items of conflict.

The job of consolidating more than 2,000 tasks, many of which were duplicates, into a single composite in terms that were meaningful to personnel in all Services proved to be an exhausting effort for all those involved. The discussions ranged over a wide variety of task writing philosophies and practices and resulted in a learning experience for all. By the end of the week-long meeting, a number of issues were resolved:

a. A single composite task list was agreed to with each Service agreeing to field validate the list prior to the next meeting;

b. since each Service would print its own job inventories, each representative agreed to modify the instructions in the final job inventory to reflect the interservice nature of the project;

c. a consensus was reached that no duty titles would be printed in the final task list, tasks would be grouped into duties and listed in alphabetical order;

d. a number of background items which would appear in identical format in each job inventory were agreed upon; and

e. a second meeting would be held to finalize the task list and to complete the development effort.

The second interservice task consolidation workshop convened, again in Washington D.C. in May 1979. Again, each Service was represented by an inventory development specialist and suject matter expert. Each Service representative reported that the task list was well received by personnel in their Service and terms were universally understandable. With only minor discussion, the final list of tasks and background questions were agreed to. The final list contained 577 task statements and each Service representative agreed to three background questions which will be identical in all job inventories - 93 duty position titles, 206 equipment items, and 37 programming languages.

Following the second consolidation workshop, the list of tasks and background questions were prepared in their final form at a single location to insure that all items emerged in an identical format. Each Service was then given a final copy from which they prepared and printed their own job inventory booklets.

THE COMPUTER PROJECT
DATA GATHERING

Each Service was responsible for selection of the survey population and the method for data collection. The Army, Navy, and Marine Corps originally proposed sampling 100 percent of their population of computer operators and programmers. The Air Force, due to the large size of the computer operator and programmer specialties - approximately twice the size of the other services combined - selected a stratified random sample of 70 percent. Table 1 presents the number of personnel to be sampled for each Service.

TABLE 1
POPULATION SIZE

USAF	3,877 (total population 5,905)
USA	2,084
USN	2,400
USMC	1,150

The data gathering processes for each Service were considerably different. The Navy and Marine Corps gathered their data by sending personnel from the occupational analysis agencies to field locations and conducting on-site administrations. In addition, these two Services mailed job inventories to those locations they could not visit. The Army and Air Force administered their job inventories totally by mail.

THE COMPUTER PROJECT
NOW

At the time of this writing each Service was gathering data and accomplishing the programming necessary to insure compatible data tapes. The project officers from each Service feel that the remaining milestones for the project - data available by 15 October 1979 and analysis complete by January 1980 - can be met. It is further anticipated that the results of the interservice analysis may be shared with the members of the Military Testing Association in 1980.

Bibliography

Interservice Training Review Organization Executive Committee Order Ten, 14 Sep 78, Ft Monroe VA

Interservice Training Review Organization 6th Annual Report, 6 Mar 79, Ft Monroe VA

Interservice Training Review Organization Procedures Manual, Oct 77

Interservice Training Review Organization Steering Committee Minutes, 21 Jun 79, Atlanta GA

Memorandum for ITRO Committee Chairman, Subject: Reorganization and Revitalization of ITRO, 4 May 79

Memorandum for Record, Subject: Interservice Task Analysis Program, 11 Nov 74, Lackland AFB TX

Memorandum of Understanding, Subject: Interservice Projects, 19 Jul 77, Lackland AFB TX

Memorandum for Record, Subject: Trip Report - Interrater Computer Operator/Programmer Survey, 25 May 79, Randolph AFB TX

Memorandum for Record, Subject: Trip Report - Task Consolidation Workshop, 26 Mar 79, Randolph AFB TX

Minutes of Interservice Occupational Task Analysis Program Subcommittee, 20 Jul 77, Lackland AFB TX

Minutes of Interservice Occupational Task Analysis Program Subcommittee, 10 Jan 79, Atlanta GA

Summary of AFSC 511XX Interservice Inventory Development, May 79, Randolph AFB TX

EFFICIENCIES RESULTING FROM THE STRUCTURED REWRITE OF OVERLAP-GROUP

Jay Charles Soper

Occupational Research Program
Industrial Engineering Department
Texas A&M University
College Station, Texas 77843

INTRODUCTION

Overlap-Group is an important part of the CODAP system of computer programs, a major job analytic tool used by many groups within the U.S. military. While redesigning and rewriting the CODAP system, Texas A&M's Occupational Research Program has made significant improvements in Overlap-Group by carefully and consistently applying the principles of STRUCTURED PROGRAMMING (hereafter abbreviated SP). These SP principles, taken together with an intelligent approach to design, testing, and documentation of software systems, may be called SOFTWARE ENGINEERING TECHNOLOGY (SET). The successful implementation of SET in the redesign of Overlap-Group demonstrates SET's usefulness to developers of medium-and large scale scientific FORTRAN programs.

WHAT DOES OVERLAP-GROUP DO?

Overlap-Group is a statistical procedure performed near the beginning of a CODAP study. It hierarchically forms groups or clusters of incumbents (workers) who perform similar jobs. Knowledge of who clustered with whom, what jobs were performed by the incumbents in the clusters, how similar clusters are to each other, and like information is vital to the job analytic process.

Because Overlap-Group considers all the tasks performed by all the incumbents in the study, it must deal with a huge amount of data. For instance, a study having 10,000 incumbents who responded to 500 task statements would have around 20 <u>million</u> bytes of raw data for Overlap-Group to analyze.

Any program dealing with that much data is expensive to use. This is especially true of Overlap-Group because of the recursive nature of the clustering process. After the two most similar incumbents are clustered, the whole process must be repeated. In a study having 10,000 incumbents, 9,999 clustering operations occur.

One further implication of the large database Overlap-Group uses is the extensive I/O time required. While quite a bit of heavy "number crunching" calculation is done by the program, that time is dwarfed by the I/O time.

To summarize, Overlap-Group has a moderately complex task to perform on a very large amount of data.

WHAT WAS OLD OVERLAP-GROUP LIKE?

Old Overlap-Group was developed by many different programmers over a period of years. Its source code revealed the fact that several different programmers with widely varying programming styles had helped in the program's evolution. Even though the program worked, there were several shortcomings in the program which necessitated a complete redesign and rewrite.

The primary motivation for rewriting Overlap-Group was the same as that for reimplementing the whole CODAP system: a desire to have a transportable system. Old Overlap-Group's authors had made full use of every extension in the various FORTRAN compilers they had used. This rendered the program useless on machines and compilers other than the one which had been used last for maintaining the program.

Other problems with old Overlap-Group included an almost total lack of documentation, sections of dead code (a "GO TO" before an unlabeled statement), and duplicated and nonfunctioning checks for various errors.

The root cause of these problems was the size and complexity of the program. Nearly 3000 lines of FORTRAN with no indentation, virtually no comments, and variable names like "X" and "JA" led to code which was difficult to understand. Whenever code is hard to understand, it is hard to modify. Many of the duplicated error checks were probably added by frustrated maintenance programmers who knew what the code needed to perform the check would look like, but couldn't find the nonfunctioning code. Code that couldn't be found couldn't be fixed, so a duplicate check was added instead.

The programmers of Overlap-Group can't really be blamed for the condition of the program. Only in the last few years has SET come to be understood. It is still not widely accepted. Undoutably the developers of old Overlap-Group were never exposed to SET.

WHAT IS GOOD SOFTWARE LIKE?

Good software in general has five characteristics; it will be USEFUL, RELIABLE, UNDERSTANDABLE, MODIFIABLE, and EFFICIENT.

USEFUL software accomplishes the task for which it was written. The task is non-trivial. In other words, it works.

RELIABLE software can be depended upon to always behave in the desired manner, no matter what is done to it. This normally means that even a completely invalid and unreasonable input data stream will not kill the program. It will end gracefully, giving the user a meaningful error message which will enable him/her to fix the problem. Furthermore, reliable software will always produce the same results when given the same data as input, no matter what other programs are in the system or what other irrelevant variables may have changed since the previous run.

UNDERSTANDABLE software is simple. The logical structure is clear to people other than the program's author. It is easy to figure out how understandable software works.

MODIFIABLE software is easy to maintain. Changes can be made with a minimum effort because the exact line of code that performs a specific function can be found quickly.

EFFICIENT software executes quickly. Needless repetition of actions is not performed in efficient software. The program does what it needs to do, and no more.

HOW DOES SP LEAD TO GOOD SOFTWARE?

Modularity is the distinguishing trait of structured programs. A top-down, hierarchical approach to designing a problem's solution will lead to good software. The concept of abstracting a problem into major phases, then solving each phase by abstracting its solution into major phases, and continuing this iterative process naturally creates a modularized solution.

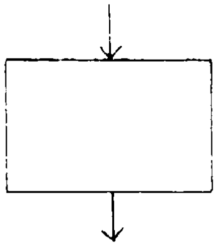
A module is a section of code which performs one specific part of the actions necessary for the total solution of the problem. Modules are organized in a hierarchy so that a high level module may call upon a low low level module to handle the details of some action. The high level module abstractly deals with that aspect of the problem, suppressing irrelevant details.

Modules are characterized by their STRENGTH and COUPLING. A strong module is one whose function is well defined, quite specific, and small for the level at which the module appears. The stronger the module is, the better. COUPLING refers to how modules interface with each other. The less connection there is, tne better. Ideally a module should be very loosely coupled, having no knowledge of how lower level modules perform their task, only knowing that they do perform it when supplied with the appropriate information. The algorithm and data structures used by a module should be hidden from higher level modules.

Structured Programming encourages modularization of programs by allowing only certain programming constructs to control the logic flow of a program. Each construct is in itself a module, having exactly one entrance and exactly one exit. This allows modules to be unplugged and replaced with minimal effects on the rest of a program.

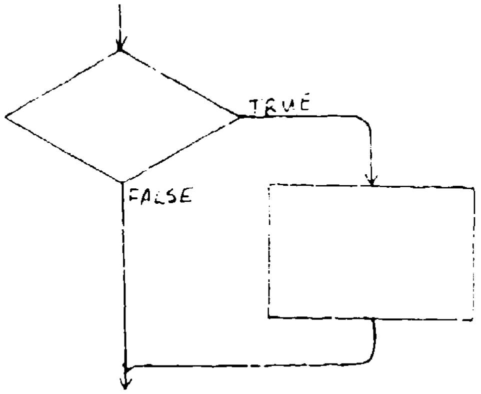
The SP constructs come in three categories: Primitive, Conditional Performance, and Looping.

The Primitive category has only one member, the simple SEQUENCE of one action after another. Obviously the conditions of modularity are met.



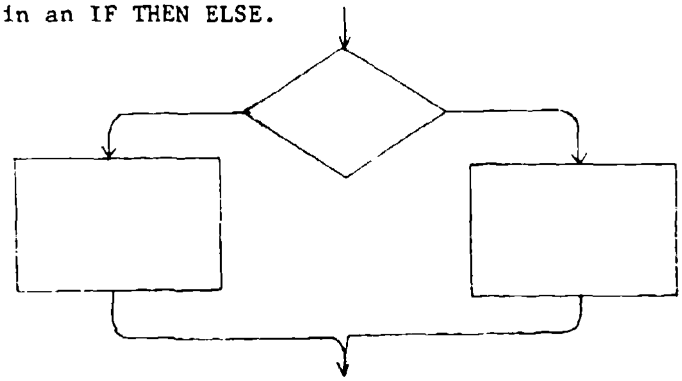
The second category of SP constructs is Conditional Performance and includes three members: IF THEN, IF THEN ELSE, and CASE STATEMENT.

An IF THEN statement causes an action to be performed only if a specified condition is met. If the condition is not true, then the action is not performed and its code is skipped.



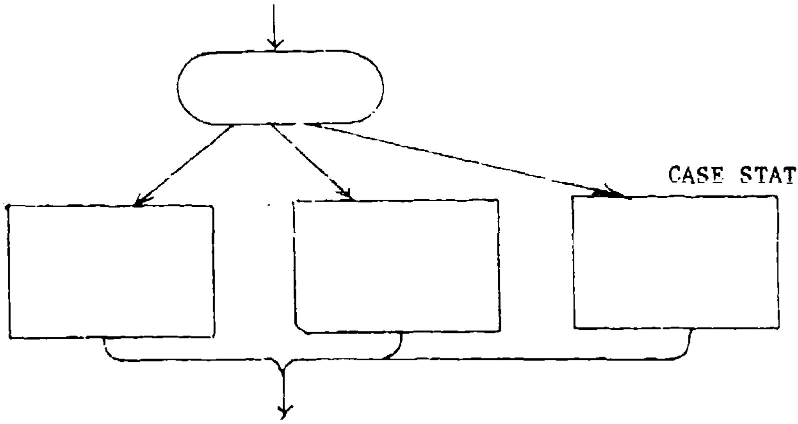
IF THEN

An IF THEN ELSE is very similar. It cause one of two possible actions to be chosen, depending on whether the condition is met. Exactly one of the actions will occur in an IF THEN ELSE.



IF THEN ELSE

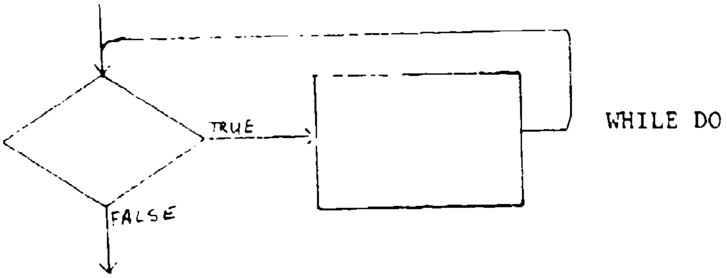
A CASE STATEMENT is just a generalization of the IF THEN ELSE to allow for more than one alternate action. Once again exactly one of the actions will be performed.



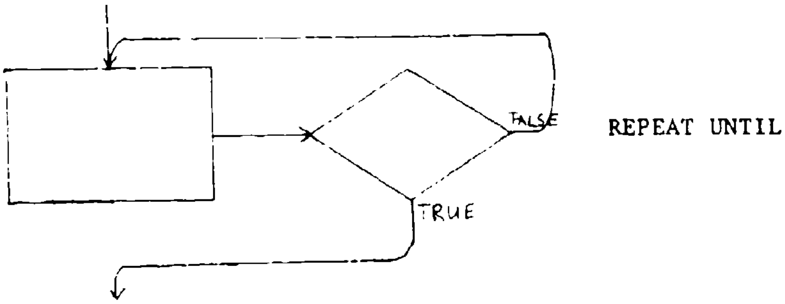
CASE STATEMENT

The third category of SP constructs also has three members; WHILE DO, REPEAT UNTIL, and FOR DO are the Looping constructs.

A WHILE DO construct causes repeated execution of an action as long as a condition remains true. The condition is checked before entry into the loop the first time, so it is possible that the action will not even be taken once.



A REPEAT UNTIL is identical to the WHILE DO except in one aspect: the loop is entered and the action is performed once before the first test of the condition.



A FOR DO loop also causes repeated execution of an action. The loop is exited when it has been executed a certain number of times, however, rather than when a logical condition becomes true. The number of times the action will occur is know upon entering the loop, while it is not known for the other two looping constructs.

The reader should note that a big box could be drawn around the various constructs which have been illustrated. If the insides of the boxes were then erased, each construct would appear to be nothing more than a SEQUENCE construct. This is how modules are viewed in SP, as black boxes which perform a function when called upon. The inner details are unimportant. The procedures are said to be "abstracted."

By restricting himself to use of only these seven constructs, the programmer greatly increases the probability that he will write good software, software that is useful, reliable, understandable, modifiable, and efficient. Furthermore, many studies have shown that use of SP increases programmer productivity. One major corporation found a ten fold increase. (They probably made a few other changes at the same time.)

HOW HAS SET BEEN USED IN NEW OVERLAP-GROUP?

Many FORTRAN programmers mistakenly believe that the SP constructs cannot be implemented in programs written in standard FORTRAN. This is not true. While a high level language designed to facilitate SP (such as PASCAL or ADA) is somewhat easier to use in writing structured code, almost any language has the necessary raw materials for building the SP constructs. SP can be done in any language having conditional go to statements and labels.

Even if the programmer must "fake" the SP constructs, he gains all the benefits of writing well structured code. Understandability, modularity, and modifiability are enhanced just as much by simulated SP constructs as they are by built-in constructs.

Several authors have suggested quick and easy ways to build the SP constructs in standard FORTRAN. Some FORTRAN compilers now have a feature to allow the programmer to write structured code and let the compiler translate it into standard FORTRAN. With a little training any FORTRAN programmer could write structured code.

In addition to the use of SP, new Overlap-Group differs from old Overlap-Group in that the new program is much more modularized. The modules of the new program are considerably smaller. In fact, the executable code fits on one or two pages for all of them. No module is larger than 3K bytes of object code.

It should be pointed out that modularity is most valuable for development and maintenance purposes. The production version of most programs can be made more efficient by copying some of the more frequently called modules directly in line with the source code of the calling modules. The modular structure will still be clear, if this is done carefully, and a significant amount of subroutine linkage overhead can be eliminated in some cases.

Another major difference between new and old Overlap-Group is the increased attention documentation has received during development. While user's manuals, maintenance manuals, and other external documents are important, the biggest change in this area is documentation internal to the programs. This involves more than just throwing in a few comments to explain what the program does.

The modules of new Overlap-Group all have several sections in addition to the executable code. These include: Name, Number, Function, Variable declarations (divided into local variables and parameters), Input from parameters, Output through parameters, Input from external data sets, Output to external data sets, Output to printer, and a Logical Outline of the program.

Executable code constitutes only a small percentage of the total source code listing. The goal in following this procedure is give exactly the right amount of information (neither too much nor too little) to make the module easy to understand and easy to modidy.

The last major difference between old and new Overlap-Group is slightly less concrete and tangible. New Overlap-Group does not use "tricks." The tasks to be performed are described in a straightforward manner, and straightforward FORTRAN is written to implement the tasks. This practice greatly increases understandability and modifiability of the program since very often the trick will work only for a particular compiler or particular computer, and the feature that makes the trick work is not generally known.

One further minor change involves the use of meaningful, mnemonic variable and subprogram names. By calling an array used as a communications region COMREG instead of JA, the understandability and modifiability of the program is enhanced significantly.

CONCLUSION

The application of SET to the rewrite of Overlap-Group has had several impacts on the program. It is much simpler, now. Because of this, it is easier to understand, modify, and maintain.

To the casual observer, the most significant difference would be the change in size of the program. While the old program approached 3000 lines of FORTRAN, plus many, many calls to assembler language subroutines, the new program is currently under 600 lines long and has now subroutines in assembler. Some work remains to be done on the new program, yet it is obvious that a significant improvement has been made.

Since some details of the database access methods remain unresolved in the new CODAP system it is not possible to directly compare execution times between old and new Overlap-Group. If the experience of the past holds true in this case, however, new Overlap-Group will run significantly faster than old Overlap-Group. This is simply because short programs take less time to execute than do long programs. Of course, new Overlap-Group has lost efficiency by using neither assembly language routines nor extensions to standard FORTRAN.

The most important feature of new Overlap-Group is its transportability. Because the program was written in 1966 standard ANSI FORTRAN and made use of no machine dependent features, it should be able to run on machines other than the one on which it was developed without changes. Since this was the original goal of the rewrite, the increased efficiency, improved understandability, and better documentation are like icing on the cake.

This project's success demonstrates the value of Software Engineering Technology, and Structured Programming in particular. Design and implementation of Overlap-Group were made easier by use of SET, and the author feels certain that maintenance of the program will be easier, also. The author hopes that his successful experience with SET will encourage other software designers and programmers to use these techniques, too.

RELATIVE TIME-SPENT SCALES - ANOTHER LOOK AT PERFORMANCE MEASURES

M. Joyce Giorgia
Air Force Human Resources Laboratory (AFHRL)
Brooks AFB Texas 78235

and

James B. Carpenter
St. Mary's University
San Antonio, Texas 78248

INTRODUCTION

The Air Force Occupational Analysis Program which was designed specifically for large scale administration and operational application, requires job incumbents to rate each task performed in their job using relative time spent ratings. Basic assumptions of the Air Force approach include, first, a comprehensive listing of all tasks which may be performed by anyone within a given specialty. Secondly, it suggests that each task is independent in terms of content from all other tasks within the listing. Given the validity of these two assumptions, the job incumbent, then, using the task level survey instrument, identifies those tasks which are part of his/her job and indicates the relative time spent on each task performed in relation to all other tasks which he/she performs. This data may then be analyzed using the Comprehensive Occupational Data Analysis Programs (CODAP), which computes the percentage time spent per task performed by each individual (individual job description) and also provides options for developing group job descriptions and group job difference descriptions.

Traditionally, major emphasis has been placed on the group job descriptions for the purpose of identification and modification of existing classification structures, personnel selection techniques, measures of job difficulty, or potential job reengineering actions. For these purposes, the accuracy of a group job description is of primary concern. Carpenter (1974) investigated the sensitivity of group job descriptions to possible inaccuracies in the individual job description upon which the group description was based. His findings suggested that descriptions comprised of more than five individuals were both acceptably stable and valid, and, as the size of the group increased, the stability of the group job description quite rapidly approached an asymptotic level, with little increase in stability occurring in groups ranging in numbers between 50 and 100. Similar findings were reported by Pass (1978) who used a correlational approach in measuring stability of group job descriptions and found the stability as measured by this technique approaching an asymptotic level at between 30 and 100 subjects. The differences in these two approaches were essentially in the use of the measuring scale. Pass used a performance-nonperformance approach, while Carpenter used the traditional relative time spent ratings. In both cases, and in other reports of similar findings (Cronbach & Gleser 1957) some form of correlational analysis was used as the measure of stability. The problems with these approaches relate to the analysis technique which is employed. It is, of course, quite possible to achieve relatively high correlations between subjects or ratings, even though the ratings themselves are consistently biased in a specific direction. Thus, the stability of the reported results may best be considered analogous to reliability; that is to say, the job descriptions in themselves are consistent as measured by the reliability coefficient, but a consistent biasing error is not precluded by the relatively high correlations.

Since stability of group job descriptions may be inferred from the above and other research, the attention of investigators concerned with the validity, in the sense of accuracy of job descriptions, must turn to attempts to indicate or specify the actual validities of the job difference descriptions, group job descriptions, and individual job descriptions. The validity of the individual job description is an inherent prerequisite to its use in job reengineering, organizational restructuring, and the development of various assignment projection models.

The historical problem with this relatively direct approach has been the establishment of accurate criterion information for use in quantifying any existing errors in the resulting job descriptions. Carpenter, Giorgia, and McFarland (1975) attempted to circumvent this problem and specifically investigate the effect of varied potentially usable scale formats on the accuracy of the derived job descriptions. Two specific approacnes to this investigation and the experimental design employed were reported in AFHRL-TR-75-63 and will now be summarized prior to considering an alternate data gathering scale historically considered as a performance-nonperformance option.

Prior Research on Relative Time Spent Scale Validity

In the first phase of the reported research, the experimental materials were five hypothetical job descriptions which specified the tasks performed and the average time in minutes spent on each task per week. Subjects for this phase of the research were 265 airmen randomly assigned to one of five groups (53 in each group). Each airman was given a job description which he/she was to assume to be an accurate representation of a job he/she was performing. Each subject then used each of five different rating scales to describe this job. The job descriptions consisted of identical task listings broken down into five major duty headings specifying the actual number of minutes per week spent on each task performed. The five job types described were developed to include a range of time-spent-on-any-one-task values of from one to five percent to one to twenty-six percent. In addition, the skewness of tasks performed varied from the normal in both directions. Each subject rated the relative time spent on each task using five different scales presented in a totally counterbalanced design.

Scale A was a 5-point relative scale with an anchor at every point. Scale B was a 7-point relative scale with an anchor at every point. Scale C was a 9-point relative scale, again anchored at each point. Scale D was a 25-point relative scale with anchorages provided for five intervals on the scale. Scale E consisted of directions to indicate the proportional amount of time spent using direct percentage estimates. A short paragraph of instructions was included with the scale definition. Since the actual time spent per week on each task was available, completely accurate percentage values could be easily computed to serve as the criterion against which the derived job descriptions could be evaluated. Rating information was analyzed using the CODAP system and individual job descriptions for each subject by scale were computed. The derived values expressed as percent time spent were then compared to the criterion values and the absolute percentage differences across all listed tasks were summed to derive a criterion comparison (CRICOM) value. This experimental measure of error in the job description was used as the dependent variable in a two-way analysis of variance with repeated measures on the scale effect variable.

The analysis revealed that the average error per task performed on the derived individual job descriptions ranged from 2.0 percent per task for job type 2 consisting of 12 tasks (positively skewed distribution) with a ratio of time spent from one to 18, to 0.7 percent per task for job type 5 which consisted of 20 tasks (relatively leptokurtic distribution) with a time spent ratio of 1 to 9. The effect of the scale utilized to derive the individual job descriptions revealed that the 5-point rating scale was significantly inferior to all other scales employed and that the 7 and 25 point scales were inferior to the direct estimate of percent time spent (treated as a relative index value) for each performed task. Of interest was the fact that the 9-point scale was either statistically equivalent or better than all other scales used in this study. Table 1 shows the average CRICOM variance across all scales employed in this study. Inspection of this table suggests a general tendency for the actual errors in the individual job descriptions to reach a relatively asymptotic level with the use of a 7-point scale. The CRICOM values were recomputed using CRICOM squared values to minimize the effects of smaller and less significant errors while maximizing the effects of major deviations in the time spent values on any specific task. Results of these analyses were, for all essential purposes, identical to those reported earlier. Thus, the first phase of the cited research shows relatively consistent findings, but does not specifically address the issue of relative scale validity, since providing the subjects with tasks performed and actual times reduces the magnitude of the potential variance and must be considered a second order approximation of the ultimate criteria.

The second phase of the cited research employed a job task inventory specifically constructed as a test instrument which included all tasks performed by Air Force basic military trainees during the six week basic training program. This inventory was prepared in accordance with the standard Air Force inventory construction methodology. Eighteen flights of basic airmen were identified as experimental subjects. Since the curriculum was highly standardized, the technical instructors who were with the flights during basic training were provided copies of the duty-task listing and requested to record, in minutes throughout each training day, the actual time spent on each performed task by the members of their flight. Validity of this criterion data was high, with the inter-rater reliability of the instructors' mean values approaching .98 (Rkk), and the single rater reliability (r_{11}) in excess of .78. With this extremely stable criterion vector available, the 834 basic trainees assigned to 18 basic training flights were randomly divided into eight groups. Each experimental group used one of 8 scales to provide the researcher with input data for the derivation of individual job descriptions. Scales A, B, and C were identical to those scales utilized in the first phase of the research. Scales E and F were similar to the 25-point and the actual percent time spent estimates as in the original research. Scale D was a 9-point scale with only the center (point 5) anchored. Scale H was a direct estimate of the total time spent in hours and minutes. Scale G was a dual composite of relative frequency times absolute-time-spent-per-performance formulation.

While Scale G was statistically superior to all other scales employed in this study, the magnitude of the actual error per task is not markedly different in terms of percentage of error. The average error per task ranged from a 1% per task for Scale G to 1.1% per task for Scale F.

The findings indicated that all methodological approaches employing relative time spent scales do, in fact, result in valid task-level time spent estimates. To further investigate the relative validity of the various scale formats, the reported data were reanalyzed applying the instructor-provided criteria to nine sets of companion flights, each consisting of two flights belonging to a single training squadron. Correlations between the individual flight criterion values ranged from .94 to .99. In this reanalysis, the 5- and 7-point rating scales were shown to be statistically inferior to the other scales employed in this study, and based on these findings and other more limited analyses, the 9-point relative time spent rating scale was recommended as providing the optimal format for use in the occupational analysis program. The 9-point scale has been routinely employed in the Air Force occupational analysis program since early 1975.

Relative Time-Spent Scales Compared to Performance Statements

The five-, seven-, and nine-point relative time spent rating scales, as well as the various other formats employed in the described research, may be considered analogous to an equal interval scaling approach in which lack of response to a task item (i.e., lack of performance) should be considered as an additional point on the scale; namely, zero. This consideration is based on the fact that all tasks within an inventory are first checked on the basis of performance-nonperformance and then rated for only those tasks which are performed in terms of the relative amount of time spent. Both historical and recent (Pass, 1978) researchers have addressed themselves to the use of the performance-nonperformance dichotomy as a rating scale for direct input to the CODAP programs. The performance-nonperformance scale would be directly amenable to CODAP analysis. In the computation of individual job descriptions, each performed task would receive an equal percent time spent value based on the total number of tasks performed by the respondent. The use of this approach would have the effect of greatly reducing the amount of time required to complete the job inventory. However, the suitability of this approach must be evaluated in terms of the inherent inaccuracies created by the restriction in response options. Since the first phase of the earlier reported research provided respondents with the specific tasks performed and the actual time per week spent in performing each of the tasks, a review of the original task respondent materials was accomplished to verify an understanding of the instructions on the part of the respondent. The data supplied by the respondents suggested that the instructions were both understood and followed, in that respondents used a relative time spent scale to record some time spent on each task performed and did in fact leave blanks, indicating nonperformance of those tasks which their job inventories indicated were not performed. It was thus possible to assign the value of 1 to each task performed and the value of 0 to those tasks not performed. This formulation then was equivalent to a performance-nonperformance dichotomy and was input to the CODAP system and individual job descriptions were derived. However, these data were not amenable to inclusion in the original ANOVA model because the performed tasks had a variance and standard deviation of zero and the inclusion of the dichotomized scale sample would have violated the assumption of equivalent variances among sample groups.

However, the magnitude of CRICOM variance was significantly greater than that reported for any of the relative scale formats and, in fact, generally approached the expected value of the variance based on an extension of the curvilinear line of best fit which could be extrapolated from the data shown in Table 1.

The second phase of the earlier research utilized a job task inventory which was administered in accordance with normal procedures, i.e., incumbents first checked each task which they performed and, following completion of this, then went back to indicate relative time spent values for each performed task. The initial check for performance of the task is equivalent to a specification of performance and all data provided by the 834 basic trainees responding to this phase of the original research were recoded 1 if performed, zero if not performed, and individual job descriptions derived using standard CODAP programs. For the group of subjects who had originally used a 5-point rating scale, the previously computed CRICOM values were compared with the CRICOM values computed from their dichotomized responses. A correlated t-test was computed across subjects and the significant result ($t = 8.74$, df=100, $p<.01$) shows the relative time spent scale to possess significantly greater inherent accuracy in terms of the subject's ability to provide an accurate portrayal of this job. Similar analyses were computed using those subjects who had originally employed the 9-point and 25-point scales. Similar significant results of a correlated t-test ($t = 9.32$, df=100, $p<.01$; $t = 9.86$, df=100, $p<.01$) were obtained. These results clearly indicate that the accuracy of the derived job descriptions is functionally related to the number of scale points available to the respondent and the consistency of the results with the original research findings for both phases suggests that the use of a performance-nonperformance response on the part of incumbents, although highly stable, may result in individual job descriptions significantly inferior to those based on scales providing a wider range of response options. Clearly, if the validity of an individual job description is considered important for personnel actions or decisions, then the scale format providing the optimal accuracy is to be desired.

SUMMARY AND DISCUSSION

The analyses reported in this paper utilized a more definitive criterion than is typically available for validating task-level time-spent ratings. Much earlier research has used various nonquantified criteria, such as respondents' evaluations of resulting job descriptions, supervisory evaluations of respondents' job descriptions, and other subjective techniques. In the original research, upon which this paper is largely based, a quantifiable stable criterion was available which could be used as a measuring stick against which the relative accuracy, in this sense equivalent to job description validity, could be evaluated. It is imperative that the distinction between the stability of job descriptions or criterion vectors, for that matter, be differentiated from the concept of validity, defined as the accuracy of the resulting job descriptions. Stability may be achieved even with relatively inaccurate job descriptions as long as the error across job descriptions stems from a stable biasing source. Thus, the high correlations evidencing stability are definitely a prerequisite and desirable characteristic. But validity, in the sense of accuracy, is imperative if individual job descriptions are to be usable in management decisions. In this sense, the original findings favoring the 9-point relative time spent scale as the optimal compromise between the conflicting requirements of ease and cost factors of administration and accuracy of the derived job description data is confirmed, as is the general tendency that respondents can in fact provide more accurate descriptions with greater response options available than generally believed. Several peripheral analyses conducted in the original research should be noted as suggestive of additional potential lines of research in this area. In a secondary analyses, the accuracy of a fully anchored 9-point rating scale, as compared to a 9-point scale anchored only at the center point, was evaluated. Results of this analysis reflect no significant difference as a function of the number of anchors provided for respondents using these two 9-point scale formulations. In his presentation at last year's MTA conference, Mitchell (1978) suggested that the number of anchors provided may, in fact, affect the accuracy of the resulting job description and suggested a need for research in this area. The earlier reported findings with regards to the relative accuracy of anchored versus nonanchored 9-point scales are in a sense contradictory to his position, but this contradiction evidences the further need for research on this topic. Phalen (1978), in his COMMENTS before the same body, outlined some pilot research to investigate the usability of derived performance ratings using frequency of performance times time spent performing. This comparison was also alluded to in the original research, since this formulation was employed as rating scale X and generally reflected relatively accurate descriptions when the results of the derived values were input to the CODAP system as a form of relative estimation. The potential incorporation and accuracy of this scale format is also suggested as a viable research activity. Finally, as a side note, there has been much conversation regarding the interrelation or potential interrelationships between the range of values which the incumbent is attempting to describe and the number of options which he/she has available to utilize in attempting to make this description. It has been assumed by some researchers that a 5-point relative scale would be optimal if the number of differentiations which the respondent is attempting to make is, in fact, on a ratio of one to five. Likewise, a 20-point scale would be optimal if the range of values is one to twenty. The research reported here does not suggest this concern to be valid. Not only were respondents able to utilize a considerably greater range of response options than anticipated, but no significant differences were observed as a function of the range of response options available versus the range of time spent in performing the tasks within the job being described. In fact, it would appear that if the number of scale options exceeded the range of values which were attempted to be described, the incumbents restricted their responses to a limited section of the available scale. Thus, the problem of spuriously increasing time spent variance as a function of increasing the number of scale points, does not appear to be significant, and the potential inaccuracy of job descriptions resulting from less scale values available than required may be significant.

Table 1 CVERALL CRICOM (ERROR) VALUES

1-pt	5-pt	7-pt	9-pt	25-pt	100-pt
46.08	17.01	15.30	14.41	15.14	13.34

OVERALL CRICOM (ERROR) VALUES SQUARED

1-pt	5-pt	7-pt	9-pt	25-pt	100-pt
303.20	63.71	50.73	46.15	47.23	37.68

REFERENCES

Carpenter, J.B. Sensitivity of group job descriptions to possible inaccuracies in individual job descriptions. AFHRL-TR-74-6, AD-778 839. Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory, March 1974.

Carpenter, J.B., Giorgia, M.J., & McFarland, B.P. Comparative analysis of the relative validity for subjective time rating scales. AFHRL-TR-75-63, AD-A017 842. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, December 1975.

Christal, R.E. Stability of consolidated job descriptions based on task inventory survey information. AFHRL-TR-71-48, AD-734 739. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, August 1971.

Cronbach, L.J., & Gleser, GC. Assessing similarity between profiles. The Psychological Bulletin, 1957, 50, 456-473.

Mitchell, J.L. Differential responses on alternately anchored job rating scales. Proceedings, 20th Annual Conference of the Military Testing Association, Oklahoma City, Oct 1978, 525-536.

Pass, J.J. Sample size and stability of task analysis inventory response scales. Proceedings, 20th Annual Conference of the MIlitary Testing Association, Oklahoma City, Oct 1978, 537-555.

Phalen, W.J. The development of a technique for using occupational survey data to construct and weight computer derived test outlines for Air Force Specialty Knowledge Tests (SKTs). Proceedings, 20th Annual Conference of the Military Testing Association, Oklahoma City, Oct 1978, 949-975.

A COMPARISON OF SAFETY TRAINING AND OTHER VOCATIONAL TRAINING REQUIREMENTS

Nancy A. Thompson and Hendrick W. Ruck

Occupation and Manpower Research Division
Air Force Human Resources Laboratory
Brooks AFB, Texas 78235

INTRODUCTION

For several years the Air Force Human Resources Laboratory (AFHRL) has been investigating many facets of Air Force training requirements to develop more efficient methods to answer the question, "What should the training content be?" At last year's Military Testing Association Convention, a paper dealing with innovations in determining training requirements was presented by Ruck, Thompson and Thomson (1978). One of these innovations, recommended training emphasis as rated by field supervisors, is now operationally collected by the Air Force Occupational Measurement Center (USAF OMC) at Randolph AFB, Texas. Also, Thompson and Ruck (1978) presented a paper at last year's MTA describing the feasibility of prioritizing job tasks in terms of hazard potential, expected frequency of accidents and other pertinent factors that could assist training designers in determining needs for safety training.

This paper will focus on the fundamental question of selection of tasks for training. To accomplish this, the problem of selection of tasks for training will be addressed from a broad instructional system design perspective as well as from a more narrow safety oriented perspective. While the broad perspective allows many possible factors to be considered in developing training requirements, it may fail to address specific requirements inherent in certain types of training. Therefore, the purpose of this paper is to compare the results of the broad instructional system design approach and the more narrow safety perspective.

APPROACH

General ratings of training priority as well as specialized ratings of hazard potential were collected from first line supervisors in three Air Force specialties: Aircraft Armament (AFS462x0), Fire Protection (AFS571X0), and Fuels (AFS631X0). Survey booklets containing the task lists for each specialty were mailed to field supervisors and ratings were obtained on the following task factors: probable consequences of inadequate performance, task delay tolerance, learning difficulty, and hazard potential. In addition, percent time spent by apprentices and journeymen performing each task, percent of apprentices and journeymen performing each task, and weighted average military grade of individuals performing each task were included as potential contributors to training requirements. The task factors that were used had been developed and field tested under several earlier research efforts (Mead, 1975; Mial & Christal, 1974; Stacy, Thompson, Thomson, 1977; Thompson & Ruck, 1978). All task factor ratings were based on a 1-9 scale.

The recommended training emphasis scale is defined as a measure of the tasks recommended for formal training emphasis (either school or OJT), based upon the ratings of 7 and 9 skill level field NCOs. The scale ranges from extremely little through extremely heavy training emphasis. The hazard potential scale was suggested by a study which evaluated human effects on nuclear system's safety (Askren, Campbell, Seifert, Hall, Johnson, Sulzen, 1976). This scale was designed to determine which tasks are more hazardous to perform than others and might, therefore, cause accidents. It is defined in terms of a variety of factors such as consequences of inadequate performance, mechanical failure, environmental conditions, etc. The scale ranges from extremely low through extremely high hazard potential.

For analysis purposes, non responses for recommended training emphasis and hazard potential were treated as zero, indicating no training recommended or no hazard potential. Tasks were treated as subjects in the study, since the properties of the tasks themselves were considered in recommending tasks for formal training. Sample sizes were 527 tasks for Aircraft Armament, 484 tasks for Fire Protection, and 374 tasks for Fuels. Analyses for this study were accomplished using Comprehensive Data Analysis Programs (CODAP) developed by Christal and Weissmuller (1976). These analyses included intercorrelations among the variables and multiple linear regression problems. The broad, instructional system design model was developed using a policy capturing approach (Christal, 1968). The criterion or dependent variable was field recommended training emphasis. For the more restricted safety model, the criterion was frequency of occurrence of accidents on tasks in the job inventory. Additionally, CODAP factor printouts (FACPRTs) were obtained to show the order of rated tasks from highest to lowest for factors in each of the specialties.

RESULTS

One of the tests used to determine the utility of safety related data was a comparison of inter-rater agreement (R_{kk}) of hazard potential ratings and the other instructional system design variables. Table 1 reports the inter-rater agreement (R_{kk}) for each task factor for each specialty. The rater agreement indices are for sample sizes of 50 raters as estimated by the Spearman Brown formula. Stable reliability estimates are reported for all task factors. The mean ratings for each variable for each specialty are listed in Table 2.

TABLE 1. Inter-rater Agreement (R_{kk}) of Task Factors for Aircraft Armament, Fire Protection, and Fuels

Task Factors	R_{kk}*		
	Aircraft Armament	Fire Protection	Fuels
Hazard Potential	.93	.98	.98
Recommended Training Emphasis	.96	.96	.96
Consequences of Inadequate Performance	.94	.90	.93
Task Delay Tolerance	.89	.96	.92
Task Difficulty	.93	.97	.95

* Rater agreement indices for a sample size of 50 raters as estimated by the Spearman Brown formula.

TABLE 2. Mean Ratings of Variables for Aircraft Armament, Fire Protection, and Fuels

Variables	Mean Ratings		
	Aircraft Armament	Fire Protection	Fuels
Hazard Potential	1.87	3.81	2.07
Recommended Training Emphasis	2.72	3.55	3.22
Consequences of Inadequate Performance	6.16	5.73	5.13
Task Delay Tolerance	4.52	3.69	3.60
Task Difficulty	4.07	5.00	4.12
Percent Members Performing	12.78	20.01	10.21
Percent Time Spent	.19	.21	.27
Weighted Average Military Grade	5.02	5.27	5.42

The second method used to test the usefulness of hazard potential ratings was a comparison of hazard with recommended formal training emphasis ratings. Table 3 shows zero order correlations between recommended training emphasis and the other variables for the three specialties. Recommended training emphasis and hazard potential ratings correlated .36 for Aircraft Armament, .52 for Fire Protection, and .30 for Fuels. While all three correlations were significant, there was a great deal of variance unaccounted for, which indicates that recommended training emphasis and hazard potential are measuring different training requirements. If each type of rating were used to establish training requirements, then two different training policies would emerge.

Of additional interest are several other univariate analyses among the variables. Table 4 shows zero order correlations between hazard potential and the other variables for the three specialties. Hazard potential correlated significantly with frequency of occurrence of accidents on tasks for Aircraft Armament and for Fuels but not for Fire Protection. Similarly, training emphasis correlated significantly with the frequency variable for all three specialties. As expected, hazard potential and consequences of inadequate performance accounted for a great deal of common variance in all three specialties. Correlations between the two were .70 for Aircraft Armament, .64 for Fire Protection, and .84 for Fuels.

TABLE 3. Zero Order Correlations Between Recommended Training Emphasis and Other Variables for Aircraft Armament, Fire Protection and Fuels

Variables	Recommended Training Emphasis		
	Aircraft Armament	Fire Protection	Fuels
Hazard Potential	.36*	.52*	.30*
Consequences of Inadequate Performance	.56*	.55*	.33*
Task Delay Tolerance	-.39*	-.75*	-.79*
Task Difficulty	.01	-.52*	-.60*
Percent Members Performing	.67*	.82*	.75*
Percent Time Spent	.57*	.69*	.53*
Weighted Average Military Grade	-.46	-.77*	-.61*
Accident Frequency	.17*	.11	.20*

* P<.025

TABLE 4. Zero Order Correlations Between Hazard Potential and Other Variables for Aircraft Armament, Fire Protection and Fuels

Variables	Hazard Potential		
	Aircraft Armament	Fire Protection	Fuels
Recommended Training Emphasis	.36*	.52*	.30*
Consequences of Inadequate Performance	.70*	.64*	.80*
Task Delay Tolerance	-.34*	-.62*	-.41*
Task Difficulty	-.04	-.02	.05
Percent Members Performing	.33*	.29*	.10
Percent Time Spent	.35*	.16*	.14*
Weighted Average Military Grade	-.70*	-.45*	-.70*
Accident Frequency	.28*	.06	.14*

* P<.025

A third analysis of the data was to determine whether hazard potential ratings added unique variance when combined with other instructional system design variables in predicting the criterion. For the broad, instructional design, the criterion of recommended training emphasis was developed. The predictors were consequences of inadequate performance, task delay tolerance, task difficulty, percent members performing, percent time spent, weighted average military grade, and squares of each of these variables. Multiple Rs for each of the specialties were .87 for Aircraft Armament, .95 for Fire Protection, and .91 for Fuels. Prediction was significantly increased when hazard and hazard squared were added to the 12 factor model for Aircraft Armament and Fuels but not when added to the equation for the Fire Protection specialty.

CODAP factor printouts (FACPRTs) were obtained to show the order of rated tasks from highest to lowest ratings for various factors in each of the specialties. Appendices A, B, and C are ordered listings of tasks for each specialty showing high positive, high negative, and minimal differences between recommended training emphasis ratings and hazard potential ratings. Tasks with high positive differences reflecting high recommended training emphasis ratings and low hazard potential ratings and tasks with high negative differences reflecting high hazard potential ratings and low recommended training emphasis ratings were identified using a cutoff score. The cutoff score was calculated using the standard error of the difference plus the difference in the means multiplied by the Z value associated with the level of significance. The cutoff was based on a significance level of .1 for a one-tailed test. Once the high emphasis low hazard and high hazard/low emphasis tasks were identified, chi square tests of significance were performed to determine trends in percent members performing, percent time spent, and task difficulty based on the two groups. For all three specialties, percent members performing was significantly higher for high emphasis/low hazard tasks than for the high hazard/low emphasis tasks (P<.01). Percent time spent was also significantly higher for the high emphasis/low hazard tasks than for the high hazard/low emphasis tasks (P<.01). For Fire Protection and Fuels, the difficulty level was significantly higher for the high hazard/low emphasis tasks (P<.01). The tasks for Aircraft Armament failed to show significant differences for task difficulty. The results of the analyses show similar trends for the Aircraft Armament and Fuels specialties. Hazard potential and training emphasis, hazard potential and accident frequency, and training emphasis and accident frequency all correlate significantly for these specialties. Prediction of training emphasis was also significantly improved for these two specialties when hazard was added to the 12 factor equation.

The results of the analysis of Fire Protection are somewhat different. Hazard potential and accident frequency as well as training emphasis and accident frequency are not significantly correlated. Also, hazard potential does not significantly add to the prediction of training emphasis. To explore possible reasons for the difference between Fire Protection and the other specialties, we reviewed the occupational survey report. (Kopala, Keeth and Lee, 1978) One of the findings was that a majority of the Fire Protection people are young, inexperienced 3- and 5- level airmen. These airmen are those who frequently perform the accident tasks. However, the more experienced supervisors made the ratings. These more experienced personnel may not perceive some of the accident tasks as potentially hazardous, and may have rated their own (supervisory) tasks as more hazardous due to the potential implications of inadequate performance. Thus, directing firefighting operations for structural fire was rated more hazardous than actually fighting the structural fires. This problem was not encountered in the other two specialties.

CONCLUSIONS

The purpose of this paper was to compare the training decisions that might be made using training emphasis recommendations reported by field supervisors with training decisions that might be made using the hazard potential of tasks. Hazard potential was of interest because it has been shown to be related to safety considerations such as accident frequency.

Hazard potential and training emphasis ratings are both reliably collected using standard rating collection techniques. The two ratings are significantly correlated; however there is considerable variance unaccounted for. Hazard potential is more closely related to the consequences of inadequate performance than is training emphasis.

In reviewing the output that a trainer would be dealing with in making training decisions, the following characteristics are apparent: tasks with high training emphasis and low hazard potential are performed by a significantly greater percentage of incumbents, take up more job-time, and are less difficult than tasks with high hazard potential and low recommended training emphasis.

It appears that the hazard potential of tasks could be an important factor in making training decisions. However, it cannot be used as the sole input for decision making, and probably should be used only in those specialties in which safety is of paramount concern. Further study of the specialties where safety is very important would help to define the differences between the broad instructional system design and the more narrow safety perspective. An important aspect of the safety perspective would be the investigation of the hazardous conditions existing when accidents occur.

REFERENCES

Askren, W.B., Campbell, W.B., Seifert, D.J., Hall, T.J., Johnson, R.C., & Sulzen, R.H. Feasibility of a computer simulation method for evaluating human effects on nuclear systems safety. AFHRL-TR-76-18. Brooks Air Force Base, TX: Advanced Systems Division, Air Force Human Resources Laboratory.

Christal, R.E. JAN: A technique for analyzing group judgment. Journal of Experimental Education, 1968, 36(4), 24-27.

Christal, R.E., & Weissmuller, J.J. New CODAP programs for analyzing task factor information. AFHRL-TR-76-3. Brooks Air Force Base, TX: Occupational and Manpower Research Division and Computational Sciences Division, Air Force Human Resources Laboratory.

Kopala, C.J., Keeth, J.B., & Lee, L.Y. Occupational Survey Report. AFPT 90-571-276. Lackland Air Force Base, TX: Occupational Survey Branch, USAF Occupational Measurement Center, 28 April 1978.

Mead, D.F. Determining training priorities for job tasks. Paper presented at the 17th Annual Conference for the Military Testing Association, U.S. Army, Indianapolis, IN, 19-19- September 1975.

Mial, R.P., & Christal, R.E. The determination of training priority for vocational tasks. Proceedings, Psychology in the Air Force Symposium. USAF Academy, April 1974, 29-33.

Ruck, H.W., Thompson, N.A., & Thomson, D.C. The collection and prediction of training emphasis ratings for curriculum development. Paper presented at 20th Annual Conference of the Military Testing Association, U.S. Coast Guard, Oklahoma City, OK, 30 Oct - 3 Nov 1978.

Stacy, W.J., Thompson, N.A. & Thomson, D.C. Occupational task factors for instructional systems development. Paper presented at 19th Annual Conference of the Military Testing Association, USAF, San Antonio, TX, 17-21 October 1977.

Thompson, N.A., & Ruck, H.W. Methods for determining safety training priorities. Paper presented at 20th Annual Conference of the Military Testing Association, U.S. Coast Guard, Oklahoma City, OK, 30 Oct - 3 Nov 1978.

Appendix A: Ordered Tasks Showing High Training Emphasis/Low Hazard Potential Tasks; Tasks with Minimal Differences Between Training Emphasis and Hazard Potential; and High Hazard/Low Training Emphasis Tasks for Aircraft Armament (462X0)

Duty	Task	Title	Seq Num	Rec Emp	Haz Pot	Diff E-H	Acc Freq	% Mem 1-48
E	144	Initiate Make Entries on, or Review Reparable Item Processing Tag Forms (AFTO Form 350)	2	5.6	.4	5.2	0	39
F	169	Inspect Suspension Gear Such as Pylons, Rails, or Racks Prior to Loading	16	6.2	2.9	3.3	0	56
G	287	Inspect Cockpit Weapons Release System Electrical or Electronic Components	32	5.5	2.5	3.0	1	27
		- - - - - - - - - - - - - - -						
M	364	Arm or Dearm Aircraft Gun Systems	312	5.5	4.9	.6	0	33
F	170	Load or Unload Non-Nuclear Munitions on Aircraft or Preload Stands or Racks	368	6.0	5.9	.1	9	57
F	162	Arm or Dearm Aircraft Armament Systems Other Than Guns	381	6.3	6.2	.1	11	62
D	123	Select or Assign Instructors	405	.1	.3	-.2	0	4
		- - - - - - - - - - - - - - -						
Q	463	Perform Munitions Transfer Procedures	478	1.6	3.4	-1.8	3	16
P	442	Perform Bombing and Gunnery Range Clearances	507	.6	3.3	-2.7	0	5
S	485	Assemble Anti-Personnel and Anti-Material Bombs	527	.3	5.4	-5.1	0	3
		X̄ (All Tasks)		2.72	1.87			12.78
		SD (All Tasks)		1.81	1.24			11.16

Appendix B: Ordered Tasks Showing High Training Emphasis/Low Hazard Potential Tasks; Tasks with Minimal Differences Between Training Emphasis and Hazard Potential; and High Hazard/Low Training Emphasis Tasks for Fire Protection (571X0)

Task	Title	Seq Num	Rec Emp	Haz Pot	Diff E-H	Acc Freq	% Mem 1-48
9	Maintain Fire Logs	1	6.3	1.7	4.6	0	52
8	Load Hoses or Make Hoseload Finishes	23	6.7	3.5	3.2	0	81
2	Drive Firefighting Vehicles	40	7.4	5.2	2.2	53	76
	- - - - - - - - - - - - - -						
13	Inspect Wet or Dry Pipe Sprinkler Systems	147	3.9	3.5	.3	0	11
18	Demonstrate Operation of Firefighting Equipment	158	5.0	4.8	.2	4	37
6	Control or Extinguish Structural Fires	181	7.3	7.2	.1	1	56
20	Rescue Personnel from Motor Vehicles	237	6.7	7.0	-.3	2	27
	- - - - - - - - - - - - - -						
1	Control Fires at Incinerator or Sanitary Fills	423	3.8	5.9	-2.1	0	20
5	Conduct Burning Pit Exercises	472	2.7	6.3	-3.6	5	8
16	Direct Missile Site Firefighting Operations	484	1.8	8.0	-6.2	0	4
	$\overline{X}$ (All Tasks)		3.55	3.81			20.01
	SD (All Tasks)		1.87	1.66			19.58

Appendix C: Ordered Tasks Showing High Training Emphasis/Low Hazard Potential Tasks;
Tasks with Minimal Differences Between Training Emphasis and Hazard Potential;
and High Hazard/Low Training Emphasis Tasks for Fuels (631X0)

Duty	Task	Title	Seq Num	Rec Emp	Haz Pot	Diff E-H	Acc Freq
F	144	Initiate Fuels Issue/Defuel Document (DOD Forms (AF Form 1994)	1	6.6	.3	6.3	0
H	256	Monitor Loses	62	4.6	1.3	3.3	0
G	200	Inspect Tanks Selected to Receive Fuels	75	5.9	2.9	3.0	12
		- - - - - - - - - - - - - -					
G	194	Fill Mobile Refueling Units from Bulk Storage	212	6.5	5.5	1.0	4
H	228	Fuel Aircraft with Pritchard Hydrant Systems	217	6.5	5.6	.9	8
C	79	Perform Quality Control Inspections	260	2.0	1.5	.5	0
H	221	Drive Tractor-Trailer Combinations	297	4.5	4.8	-.3	1
		- - - - - - - - - - - - - -					
H	234	Fuel or Defuel Aircraft with R-2 Heil Tank Trucks	328	3.7	5.6	-1.9	1
J	292	Transfer Liquid Breathing Oxygen to Aircraft	365	1.6	5.3	-3.7	0
I	342	Issue Hydrazine	374	1.6	6.3	-4.7	0
		X (All Tasks)		3.22	2.07		
		SD (All Tasks)		1.92	2.05		

THE NEW CODAP SYSTEM -- DESIGN CONCEPTS AND CAPABILITY

Richard W. Dickinson

Occupational Research Program
Industrial Engineering Department
Texas A&M University
College Station, Texas 77843

INTRODUCTION

The CODAP system is presently being used by a number of agencies in the Department of Defense to conduct basic and applied job analytic studies. CODAP is an acronym for Comprehensive Occupational Data Analysis Programs and consists of over 40 specially designed computer programs for analyzing occupational-personnel data. At present, the CODAP system is designed and written to operate on specific types of computers. This situation has developed to the point that several CODAP systems exist, with each one dependent on a specific computer and thus incompatable with each other.

The fact that CODAP is machine dependent could effectively isolate some potential users from benefitting from the convenience and flexability these programs offer in the handling of occupational-personnel data. CODAP's machine dependent nature makes distribution of the system contingent not only on an agency's need for such a job analytic tool but also on the agency's computer hardware. CODAP should be available to any agency that can benefit from this unique tool.

Recognition of the machine dependent nature of CODAP and the inherent distribution limitations such a situation imposed prompted the staff of the Occupational Research Program (ORP), Industrial Engineering Department, Texas A&M University to contract with the Office of Naval Research and Naval Occupational Analysis Center (NODAC) to rewrite the export version of the CODAP system (designed to run on IBM and related equipment) in ANSI (American National Standards Institute) FORTRAN. It was felt that such a rewrite would produce CODAP system programs that would be relatively independent of machine design, and thus would allow a significantly larger group of agencies access to this job analytic tool.

Initially, a progran by program rewrite of the CODAP system was envisioned. As the project progressed though, it became increasingly obvious that the best interests of job analysis would be served by a complete top-down redesign of the CODAP system. Although it is true that a program by program rewrite of the CODAP system into ANSI FORTRAN would make the system more transportable, it was the opinion of the staff at ORP that such an approach, while laudable, would not effectively deal with the system's major problems; that of lack of flexibility and the restrictions on data base access.

As things stand now, CODAP consists of many main programs and subroutines, each with a specific function. As job analysis evolved and new capabilities were required, more and more computer programs were written and added to the system, resulting in a huge computer system with little inherent flexibility. Compounding the problem, changes to a part of the system may have adverse effects on the rest of the system requiring much time and effort to ensure compatability. Even relatively minor changes or additions require extensive knowledge of the system and, given the poor state of internal system documentation, such expertise is difficult to acquire.

Presently, various restrictions exist in the CODAP system pertaining to the way information in the data base can be accessed. Information relating to the history or background of incumbents is stored and accessed in a different fashion than information collected on the tasks the incumbents perform. This situation has resulted in two basic sets of programs in the CODAP system. One set of programs were designed and written to access and report on history information, and another set of programs were written to report on task information. Any comparison of data vectors reported by the two basic sets of programs require extensive modification of existing computer code or requires completely new code to be written. This point emphasizes the fact that the system is inflexible in its ability to allow communication between various programs.

Consideration of the limiting aspects of the present CODAP system, along with the evolving needs of job analytic research, has led the staff of ORP to begin development of an integrated data management system that will be flexible and easy to use. In this system, all data will be equally accessible. The user will communicate with the system through the use of a simple, English-like language specified in free format (not restricted to specific card columns). In a single run, the user will be able to answer many questions -- thereby reducing computer overhead. It will be possible to conditionally create, modify and report data vectors -- with any data vectors reported becomming data themselves for further processing. The system will be written to enhance transportability, allowing many agencies to have access to this method of job analysis.

THE CODAP LANGUAGE

One of the initial specifications requested of any redesign of the CODAP system was that communication from the user to the system be in the form of English language commands. At present, input to existing CODAP systems consists of control cards in which computational options are chosen by user indication in specific card column locations. For example, to specify that all computations are to be performed on adjusted ratings in the REXALL program of the IBM export version of the CODAP system, the user is required to place a "1" in card column 17 of the first control card. Such requirements increase the probability of errors, hinder memorization of system operation and force the user to spend an inordinate amount of time concentrating on trivial aspects of computer communication. For these reasons, specifications were developed for system imput to be defined through the use of an easy to learn English-like language in free format.

At the heart of the new CODAP system is the interpreter. The interpreter is broken-up into two phases: the syntax analysis phase and the execution phase. The syntax analysis phase is concerned with reading in the CODAP source language (this is the English-like language the user will submit to the system), analyzing the statements for syntactical or logical errors, converting the English-like keywords of the language to a numerical representation for use by the system and, if no errors have been found, transferring control to the execution phase of the interpreter. The execution phase of the interpreter will then perform the operations specified by the syntax analyzer.

Like any language, the new CODAP system source language has its own vocabulary and syntax - words and the rules for putting them together. Generally, CODAP source statements consist of an identification of the operation to be performed on the data base, an indication of what information in the data bse is to be processed, and additional keywords that specifically describe details of the operation being performed. Users should find the keywords of the CODAP language easy to remember as they were chosen to reflect their function. The rules and syntax of the CODAP language will be outlined in the users' guide that will be supplied with the system as part of the overall documentation.

THE CONCEPTUAL DATA BASE

Task inventory data in the new CODAP system can be thought of as a two dimensional matrix, with the rows in the matrix representing all the data for a variable and the columns of the matrix representing all the data for an incumbent (see Figure 1). Through the use of the SELECT procedure, the user may identify and label aggregates of columns (referred to as "Groups") or aggregates of rows (referred to as "Modules") for use in later operations.

To illustrate,the following CODAP language source statement:

```
SELECT FROM COLUMNS, G1 (I1, I2, I3),
     'REMARK ASSOCIATED WITH THE COLUMN SELECTION VECTOR G1'.
```

serves to identify the columns of interest (I1, I2 and I3), labels these columns "G1" and stores this selection vector for reference at a later time.

The above SELECT statement represents the simplest case for identifying columns of interest. Through the use of the SELECT procedure, very complex relationships may be tested by the user as criterion for column identification.

Once an aggregate of columns or rows has been identified and labeled, the user may then restrict processing to those rows or columns by referring to the label in another procedure. For example, the following CODAP language source statement:

```
CREATE ROW FOR G1,
     BOB:=H1+H2, 'REMARK ASSOCIATED WITH BOB'.
```

creates a new row called "BOB" (consisting of the rows "H1" and "H2" added together), but restricts processing to only those columns identified by the label "G1" (I1, I2 and I3). As a result of the execution of the above CREATE statement, "BOB" would be added to the data base as a new row and would have a length of three elements (see Figure 2).

The Concept of Symmetry

One of the foremost design concepts incorporated into the new CODAP system is that of symmetric access and retrieval of information in the data base. In its most general sense, the term "symmetric" implies that any operation performed on rows in the data can also be performed on columns. In a previous example, the statement:

```
SELECT FROM COLUMNS, G1 (I1, I2, I3),
     'REMARK ASSOCIATED WITH THE COLUMN SELECTION VECTOR G1'.
```

served to identify the columns I1, I2 and I3 (refer to Figure 1) with the label "G1". Conversely, the user could just as easily have specified:

```
SELECT FROM ROWS, MOD1 (T1, T2),
     'REMARK ASSOCIATED WITH THE ROW SELECTION VECTOR MOD1'.
```

which would serve to identify the rows T1 and T2 with the label "MOD1". The user could then specify this label to restrict future processing to the rows identified. Following on this, the statement:

```
CREATE COLUMN FOR MOD1,
     SAM:=I1+I2, 'REMARK ASSOCIATED WITH SAM'.
```

creates a new column variable (SAM), consisting of the product of columns I1 and I2, and restricts processing to the rows identified by the label "MOD1" (T1 and T2).

In the above described fashion, the user may identify either rows or columns of interest, label the rows or columns, and then use this label in a procedure to alert the computer system to the direction of the operation being processed. A conceptual representation of the data base following these symmetric operations is found in Figure 2. A requirement for the successful symmetric access and retrieval of information is that the user must always conceptualize the data base as a two dimensional matrix consisting of rows and columns (see Figure 1). Although it is true that the data will not really be stored in the form of a matrix, the computer system will be acting as if it were in this form and will assume that user commands are a function of this.

Efficiency Considerations of Symmetry

The speed at which information can be retrieved from the CODAP data base will depend on the direction of the operation requested by the user. If data is being summarized on variables across incumbents (on rows across columns) execution should be very fast. On the other hand, if data is being summarized on incumbents across variables (on columns across rows) execution will be considerably slower, but not prohibitively so. This discrepancy in execution speed as a function of access direction is due to the way the data will be stored in the new CODAP system. Since most data summarization will be on variables across incumbents, the data storage method chosen should reduce the number of I/O (Input/Output) operations to a minimum.

DOCUMENTATION STANDARDS

One of the most important aspects of computer programming is the care taken to ensure that the function and operation of products produced is fully explained and documented. Documentation in the new CODAP system will consist of three types: internal, external and user.

Internal Documentation

Some of the most useful documentation is that which occurs within the program source code. Well written comments can greatly increase the ease of understanding the operational details of the code and can act as an aid to programming. Every unit of code (the term "unit" refers to a set of code that performs a specific operation) should be commented in such a way as to explain its function and the general procedure followed to accomplish it.

Variables within the code should have names that reflect their content and function. Using mnemonically significant names for variables and subroutines will greatly enhance understanding the function and flow of code.

External Documentation

A systems maintenance manual will accompany the new CODAP system describing basic philosophy, data structures and how the different aspects of the system will operate together as a function of design. There will be a HIPO (Input-Process-Output block diagram) package for the system along with a prose description of inputs, outputs, purpose, error handling and other relevant information to ease understanding. There will also be a hierarchy chart showing driver - subroutine relationships among units of code.

User Documentation

In a system that potentially could contain as much information visualized for the CODAP system, procedures must be designed into the system to allow user self-documentation. In the new CODAP system any time information is added to the data base (e.g., when the user creates a variable, or produces a vector for storage from a procedure) the user is required to supply some descriptive remarks associated with the information that will be stored for later reference. If the user should desire to examine what information exists on the data base, a procedure will be available in the system to list out specific contents, producing a report indicating the name of the information, the length and the remark that was stored with it.

User documentation will also consist of a Users' Guide. This document will describe the operation of the system to the user, and will be liberally supplied with examples to fully explain the conceptual framework of the data base and how the CODAP language is to be written to generate desired reports.

FIGURE 1

Conceptual Representation of New CODAP Data Base

Incumbents (Columns)

Variables (Rows)

	I1	I2	I3	I4	I5	I6
H1	2	1	1	3	6	4
H2	22	37	31	28	46	50
T1	25	0	15	25	0	25
T2	50	100	15	25	50	0
T3	25	0	30	25	0	50
T4	0	0	40	25	50	25

FIGURE 2

Conceptual Representation of New CODAP Data
Base Following Symmetric Operations

	I1	I2	I3	I4	I5	I6	SAM
H1	2	1	1	3	6	4	
H2	22	37	31	28	46	50	
T1	25	0	15	25	0	25	25
T2	50	100	15	25	50	0	150
T3	25	0	30	25	0	50	
T4	0	0	40	25	50	25	
BOB	24	38	32				

SELECTION VECTOR G1

I1	I2	I3

SELECTION VECTOR MOD1

T1	T2

SOFT SKILL ANALYSIS IN THE ARMY

LTC Bradford L. Walton
Chief, Occupational Research & Analysis Division

Training Developments Institute, HQ, TRADOC
Fort Monroe, VA 23651

INTRODUCTION

To betfer understand how this problem is being addressed by the Army and by whom, my parent organization and office should be explained so as to better understand our interface with the Army's training system. The Training Developments Institute is an agency within US Army Training and Doctrine Command located at Fort Monroe, VA. TDI was organized in 1975 to develop effective, efficient and job relevant training through application of proven training technology in the US Army's service schools, training centers, and NCOA academies; to train TRADOC school and training center staff and faculty in the application of proven training technology in the development, administration and evaluation of individual training and to formulate training policy in the Army.

TDI was organized to:

Assist TRADOC schools and Army training centers in the development, operation and evaluation of job-relevant, criterion referenced, performance oriented individualized training systems and to serve as a catalyst to assist service schools and training centers to institutionalize the systems approach to training.

Search for improved training techniques and training development practices which will allow more efficient and more effective individual training in the US Army.

To understand my division (Occupational Research & Analysis Division) you should know that it was organized in 1977 to meet the needs of TRADOC in the area of occupational analysis/job and task analysis (cognitive and psychomotor skills), supporting TRADOC's 23 service schools which are responsible for ± 400 enlisted and officer career fields. The mission of the Occupational Research & Analysis Division is to research, analyze, design, develop, implement and evaluate TRADOC policy on job and task analysis. Included functions are the development of regulatory and procedural guidance, supporting training materials, Job & Task Analysis terminology, act as the proponent agent for this field within TRADOC, and as the Army's representative in this area at all professional and state-of-the-art meetings with academia, Department of Defense and/or English speaking countries.

My major activities are:

Research - Extensive research efforts to support the ORAD mission.

Training Materials - Development of training materials to support ORAD mission.

Job and Task Analysis Proponency - Act for the Deputy Chief of Staff for Training, TRADOC in this field (inter- and intra- service and international).

Documentation Proponency - Develop and publish TRADOC documentation of job and task analysis (regulatory, procedural, definitions and training).

Evaluation - Provide job and task analysis expertise to TRADOC to evaluate on-going job and task analysis procedures and utilize feedback to refine on-going processes.

Professional Interface - Maintain on-going dialogue with agencies/activities involved in job & task analysis.

Within the context of my mission I have the following documentation proponency:

TRADOC Regulation 351-4, Job & Task Analysis

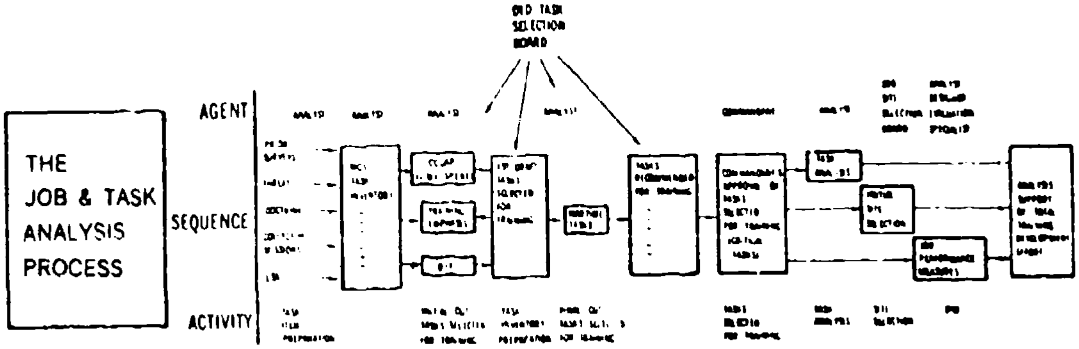
TRADOC Regulation 350-2, Development, Implementation and Evaluation of Individual Training

TRADOC Pamphlet 351-4, Job & Task Analysis Handbook

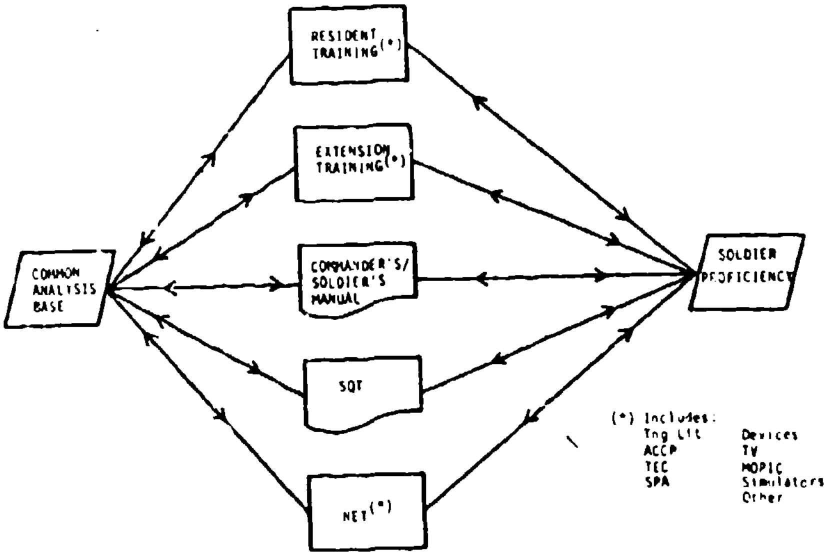
TRADOC Circular 350-3, Individual/Collective Training and Development Glossary

TRADOC Job and Task Analysis Course (27 self-paced modules - continuing activity with + 25 additional modules being completed within next 12 months.

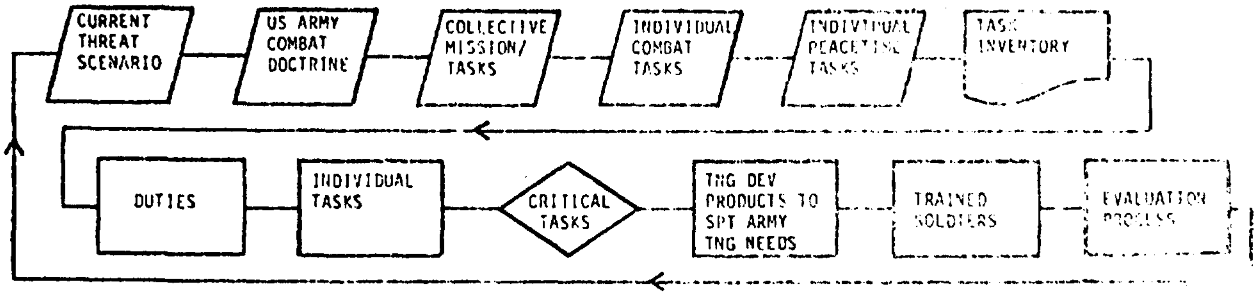
Through our efforts the Army's Job & Task Analysis Process



was developed and incorporated into the documents previously addressed. This process has a fundamental requirement: that all training developments products be based upon a common analysis data base.



This data base must be premised on a complete analysis of Army doctrine, threat, mission, collective unit tasks and new/present equipment/systems.



To proliferate this philosophy and the regulatory and procedural guidance to support it TDI/ORAD developed the documents addressed above. These are now in the hands of our analysts in the schools. One of the voids that we discovered in our research, however, was how to approach the area of cognitive skills in a systematic manner and one which could be readily taught to our analysts (Note: the average military analyst is an E6 or 03 and on the job for a period not to exceed 12 months, and one that rarely has any experience or training to support this position). To this end we have expended considerable effort in trying to resolve this issue.

The subject of "Soft Skills" (i.e., those skills requiring a cognitive vs physical ability has plagued the US Army training developers for many years. Many interim measures have been taken to design training programs however these "short-stop" measures were not systematically developed and in most cases were created (badly) to address a perceived training void. This, at a time when limited capability existed on needs analysis/assessment in the TRADOC community. The resultant factor from this ignorance was less than adequate training and testing programs for the "fuzzy" (soft skill) tasks.

With this brief introduction to the problem I am sure the reader (who I presume to be a training developer) can readily relate with the problem. Many have encountered this vast gray area that continues to be very elusive and most difficult to delineate the procedural approach needed to analyze these tasks let alone provide substantial data on which our training designers can properly design a training package. Up to this point in time we (TRADOC) found many practioners with "the answer to our problem." In reality they knew less than we did but tried to package their product more creatively. Needless to say this did little to meet our immediate needs.

As a result of this interface the Occupational Research and Analysis Division embarked upon a research effort to assist the TRADOC training community. Through our in-house research we learned considerable: in fact we learned so much that we agreed we knew very little.

What is a soft skill? A soft task? As you can imagine consensus was never reached on a definition for these words. Let us review our standby expert for definitions - Webster's Dictionary. Soft (per Webster) is defined as something "...underly susceptible to influence." Task is defined as "... the ability to use one's knowledge effectively and readily in execution or performance ... a learned power of doing something competently: a developed aptitude or ability ...". In the Dictionary of Education, skill is further defined as "...anything that the individual has learned to do with ease and precision; may be either a physical or a mental performance."

In December 1972 the Army's Continental Army Command (CONARC) the predecessor of TRADOC and FORSCOM, conducted a conference on soft skills (held at Fort Bliss, Texas). Soft skills were, at that time, referred to as "...Job related skills involving actions affecting primarily people and paper, e.g., inspecting troops, supervising office personnel, conducting studies, preparing maintenance reports, preparing efficiency reports, designing bridge structures."

In a presentation by HumRRO at this conference soft skills were addressed as being: "...(a) Important job related skills, (b) which involve little or no interaction with machines (including standardized forms) and (c) whose application in the job is quite generalized because the situation or context contains a great deal of uncertainty (i.e., we don't know much about the physical and social environments in which the skill occurs and we don't know much about the consequences of different ways of accomplishing the job function). In other words, those job functions about which we know a good deal are hard skills and those about which we know very little are soft skills."

A recommendation made by this conference, which was lost in the interim years (72-79) was that the "...use of the terms 'soft skill' and 'hard skill' be deemphasized or discontinued" because of "...their vagueness." It apparently had little impact since the recommendation was never acted upon.

As mentioned, perceptions as to what a soft skill/task is varies by person. To understand how these terms were perceived by our TRADOC Service Schools, each was quieried for a response. As suspected, the range of responses was dramatic. Notional examples are at Table 1 below:

Transportation School: Those practices that are not clearly identifiable as to either a definite beginning and end, or initiating cues for performance.	Admin Center: Some of the difficulty in analysis of soft skill tasks is perceived rather than real....the only problem....level of specificity....a second misperception is that all mental tasks are soft skills and defy analysis... as long as...(it) can be specified and real world conditions and standards specified the task can be analyzed.
AG School: Essential qualities of a softs skill- o They involve processes that span time o Are discretionary in several senses o Accomplishment is based primarily on a wide range of knowledge o Human interaction and behavior play import roles o More than one possible solution or course of action may be successful, or at least be acceptable	Intelligence Center: Standards may well be the governing factor soft vs hard skill determination. Personal perception of standards appears to be one criteria, value judgement is another. The very lack of continuity or consistency of standards for soft skill areas may well be the focus of study.

TABLE 1
School Perceptions:Soft Skills

Prior to and concurrent with this action ORAD was involved in a phased operation to pursue this problem area in more detail. This included the action above (Phase I) and the additional phases as delineated below:

PHASE	EVENT	SCHEDULE
Phase I	School Perceptions on Soft Skill Analysis	Completed (Jun 79)
Phase II	Soft Skill Analysis Symposium	Completed (23-26 Jul 79)
Phase III	Soft Skill Analysis Seminar	Completed (14-16 Aug 79)
Phase IVa	Draft Soft Skill Analysis Chapter for TRADOC Pam 351-4	1st Qtr, FY 80
Phase IVb	Initiate Strawman Soft Skill Analysis Training Modules	1st Qtr, FY 80
Phase Va	Field Review/Comments	1st/2d Qtr, FY 80
Phase Vb	Soft Skill Analysis Process Validation	FY 80
Phase Vc	Training Module Refinements	FY 80
Phase VI	Final Publication of Chap 9, TRADOC Pam 351-4	4th Qtr, FY 80

The strategy involved in this approach was to obtain assistance from the Army Research Office (ARO), Durham, NC, through the Scientific Services Program (SSP) in the form of scientists whose area of expertise was not available within the Army. Support of the five personnel listed below was obtained from ARO during the summer of 1979:

Dr. Robert K. Branson, Florida State University
(Project Coordinator)
Dr. Charles Reigeluth, Syracuse University
Mr. Ivan Horabin, Independent Consultant
Dr. Robert Gange, Florida State University
Dr. David Merrill, Courseware Inc.

At the Soft Skill Analysis Symposium (Phase II) held at the Judge Advocate General's School, Charlottesville, VA, the five SSPs above plus selected representatives from TDI, Army Research Institute (ARI), TRADOC schools and USA Management Engineering Training Activity (AMETA) participated in a sharing of knowledge, experience, current problems, and previous lessons learned with the resultant factor being the development of a model to serve as a base on which our efforts can be supported. Although refinements are anticiapted/expected, its methodology will cause no disruption to the current process in use within TRADOC and in effect, becomes but an extension of our present task analysis process and becomes a natural continuation of this process and addresses a process previously unidentified and/or properly defined in the Army. As such we coined this process as an "extended analysis" to properly reflect the process required.

The Seminar (Phase III) was held in Hampton, VA. Attendees included representatives from all TRADOC service schools (officer and enlisted anlaysis representatives) as well as the Army Health Services Command, Judge Advocate General's School, Army Research Institute, Defense Language Institute, Defense Information Center & School, HumRRO, USA MILPERCEN, the USAF Occupational Measurement Center, and the SSPs noted above minus Gagne and Merrill). Its purpose was to expose all the TRADOC service schools to this model, solicit feedback and elicit more examples/problem areas to negate our attempt to impose a process that would not be conducive to our operations or be restricted at the onset by ideosyncrasies within a school that had not previously been identified.

The result of this seminar was a fruitful exchange of information and proved very useful in our research endeavor. As such, guidance is being prepared for our schools for the extended analysis process. This will take the form of a chapter in our Job & Task Analysis Handbook and self-pacing modules for our Job & Task Analysis Course.

Essentially, our model will be a process that:

- o provides a means to "extend" our analysis process (which is presently defined)

- o uses "transfer tasks" as the vehicle to convey the cognitive task

- o will support the Army's officer and enlisted career management strucures.

- o will not disrupt the TRADOC service school current process and use of ISD.

To complement this introduction to the Army's effort on "soft skill analysis" I will be followed by reprsentatives of the Army's Military Police School, Fort McClellan, Alabama, the Judge Advocate General's School, Charlottesville, VA, my staff and one of our scientific support personnel previously addressed. These individuals will highlight how the analysis process is currently being accomplished and more on our process and its implementation strategy.

SOFT SKILL ANALYSIS IN THE MILITARY POLICE CORPS

Mr. Fred Casey
Military Police School

The US Army Military Police School is responsible for conducting Law Enforcement Training for a varied target audience. Courses of Instruction in Law Enforcement are designed to meet the needs of the individual soldier entering the Law Enforcement field as well as the NCO or Officer with extensive training and experience in the field.

One of the most challenging courses for the Military Police School to design and develop is the one that trains the Army's Military Occupational Specialty (MOS) 95B, Military Policeman. This course must train a recently inducted soldier in both the basic and advanced individual skills required to perform as a military policeman. The Military Police School has recognized for some time that this training is heavily oriented to "Soft Skills." In efforts to accomplish the task analysis for the Job (Military policeman) it became evident that the majority of the tasks requied by the job were not procedural tasks but rather highly non-proceduralized tasks containing an infinite number of variables. These variables existed in the following components of the task:

- Cue
- Conditions
- Standards

For example, one of the more difficult training problems for the training designer/developer is to identify the range of cues. That may signal when a law enforcement task is to be performed - to train the military policeman to recognize when to perform the task is very critical. Realizing that there may be an infinite number of cues for a task such as "Determine probable cause for apprehension." Recognizing the variables involved in soft skills and determining how to account for the variables during an attempt to perform task analysis is a complex problem.

As stated earlier, variables exist in the cues, conditions, and standards of most law enforcement "soft skills" (examples above). The problem the analyst and designer must solve is -- "How do we identify and record the most prominent of the variables and eventually construct a "consensus" task containing a representative cue, condition and standard? One tecnhique that our analysts have used with some degree of success is detailling the task in an algorithm (see Incl 1). The construction of the algorithm assists the analyst/designer in exploring a large number of the variables that exist within the soft task. In addition, the algorithm has proven to be a good source document for making design/development decisions relating to method, media, and testing strategy.

The "soft skill" tasks inherent to the military policeman's job (MOS 95B) possess several common characteristics. They are:

- Tasks are judgement intensive.
- Tasks involve interpersonal communications.
- Tasks have legal implications.

These characteristics of tasks make them extremely difficult, if not impossible, to proceduralize. This has prompted the Military Plice School training developers to seek some alternative for performing task analysis in the soft skill areas. The use of the algorithm, as an aid, has been helpful to our training developers in this effort, but is terribly inadequate to accomplish the objective of "analyzing soft skills."

SOFT SKILL ANALYSIS
AT THE JUDGE ADVOCATE GENERAL'S SCHOOL

PETER K. PLAUT
Major JAGC
Chief, Nonresident Instruction

The present concern over soft skill training development is not new to the Army. Nor is successful training of soft skills. The military has trained leadership, management, counseling and other soft skills for some time. The current effort does, however, represent a serious commitment to getting a handle on the analysis phase of soft skill training.

One of the institutions traditionally concerned with soft skill training is the Army's law school, The Judge Advocate General's School. That school provides professional education for Army lawyers and serves as proponent for legal subject instruction throughout the Army. The School has been a sounding board for the problems, ideas and models that have been developed in the TRADOC examination of soft skill training analysis.

The training development process at The Judge Advocate General's School has not involved the formalized, separately identifiable analysis seen in the larger Training and Doctrine Command schools. The JAG School has the advantages of being relatively small and of developing training in just a few areas of Army-wide interest. This makes it possible for one individual to see the entire cycle from analysis through evaluation of training, and permits abbreviation of the analysis process without sacrificing quality of the training product. At first glance a casual observer might conclude there was too much in the way of abbreviation. However, once one gets beyond the superficial differences of terminology and organization, the same essential steps used throughout the Army for instructional systems development can be found in the design of legal subject training for the soldier.

The topic of interest is not overall training development, however, but rather the analysis process for soft skills. In anv discussion of soft skills, there seems to be considerable confusion over what that animal is. Soft skills seem nebulous and difficult to grasp. In an attempt to clear up some of the fog, let us refine what is meant by the term soft skill.

From the perspective of the work done at the JAG School, there are several essential qualities of a soft skill. There may be other aspects, and perhaps the qualities noted below might be expressed differently. However, soft skills appear to have these common characteristics:

They involve processes that span time. This is in contrast to a mechanical task, which, although it could take a lot of time, focuses on a definite set of steps and a definite point of accomplishment. Soft skill tasks are often prospective and preventive. They involve concepts such as planning, analysis and development. The doer must draw on past experience and events, apply knowledge which extends beyond mechanical skills, and project action into the future.

Soft skills are discretionary in several senses: the doer must determine whether any action is required and must decide on the timing. Most critically, the soft skill function involves varying the conditions of performance. In other words, the doer of a soft skill can establish and change the conditions of performance.

Accomplishment of a soft skill is based primarily on a wide range of knowledge rather than mechanical skills. This is an area where rehearsal of steps and learning a process is not the entire answer. Viewed from another angle, it is difficult to isolate specific knowledges required to accomplish a soft skill task.

Human interaction and behavior play important roles in soft skills. Human capabilities, reactions and emotions must be evaluated. The collective personality of a group of people is often involved.

Soft skills frequently concern matters where more than one possible solution or course of action would be successful, or at least be acceptable performance.

There is another quality of soft skills that is more definitional than analytical: a soft skill task defies precise identification of procedural steps or measurement of job proficiency.

These qualities, especially the last concerning measurement of job proficiency, pose problems for the training analyst. The basic concept of a systems approach to instructional development is that once job proficiency is defined in measurable terms, then job skills can be taught and post-training proficiency measured in job specific areas. The difficulty is obvious: as a category, soft skills may not fit the ISD model.

This difficulty may be more semantic than real. It's all a matter of how you divide up the ISD process. To illustrate, let us look at an example from legal training.

A foundation of any military system is discipline. Instilling and enforcing discipline are functions of command. The legal aspects of enforcing discipline are items of training interest. Enforcing discipline is too broad a term for training analysis, however, and is more a responsibility than a job or task. But it is a departure point for analysis. Several legal functions can be identified for purposes of an initial job and task analysis plan. Examples at this juncture include:

--Conduct investigations concerning suspected criminal offenses.

--Administer the military court-martial system.

--Make searches and seizures.

--Administer nonjudicial punishment under Article 15, UCMJ.

These statements are sufficiently detailed to initiate a job analysis. Perhaps a better term would be ducy or function analysis. In any case, the objective is to find out what positions and grades are involved with some phase of the task. Such analysis also refines task statements. Search and seizure, for example, breaks out to:

Gather information from all sources concerning possible criminal misconduct.

Evaluate the information to determine whether a search should be conducted.

Decide what role to play in the search process.

Relay information and make recommendations.

Determine the proper legal basis for the search.

Determine if probable cause exists.

Authorize a search.

Conduct a search.

Safeguard evidence.

This list shows the major tasks involved in search and seizure. If you think about it, these are really duties. This is important, as will be explained later.

For the analyst, the desire is to determine where these various functions are performed, and who in the military organization does them. The analysis effort is essentially the same as for hard skills. The only difference is that here the approach is by topic and not initially by job or duty position.

An analyst can gather this sort of information for the other topics mentioned earlier. Very quickly the process develops extensive lists of tasks for each topic. These in turn can be broken out and refined for each grade and duty position. What that leads to, however, is a nearly unmanageable collection of lists.

It is helpful here to step back and consider that the overall objective is effective training. Analysis to refine tasks may not lead to such training. The problem is that further task analysis leads only to overly situation-oriented data that does not help in what the analyst is supposed to do.

In the traditional Instructional Systems Development sequence, analysis runs through development of detailed task lists broken out by position and grade. Analysis shows who does what, and how they do it. This serves as the basis for selecting tasks for training. Analysis ends with the decision of what to train.

In the hard skill arena, this is a convenient division between analysis and training development. For soft skills, the break may come earlier. At least for legal training, the analysis phase extends only to the duty level. It is here that functions are pegged to certain ranks and duty positions. Demographic data and information concerning required skills and knowledges is documented at this level; and it is here that duties are selected for training. The details of how tasks are performed are not developed in the analysis phase.

This duty level decision marks the transition from analysis to training development. It recognizes the essential characteristics of soft skills, as noted earlier. These skills involve too many variables of discretion, human interaction, passage of time and so on for detailed task analysis to help at all in selecting tasks for training. Additional work by the analyst on how the job is performed is not necessary. The training developer will do the same work later.

If you look at soft skills from the entire panorama of training development, it is not worthwhile for analysis to detail all the steps involved in some function such as search and seizure. It can be done but you wind up with an enormous algorithm that covers only one contingency. This only confuses the process of selecting tasks for training.

Obviously, somebody has to determine how a task is performed. For soft skills, the work should be done by the training developer. The analyst would be unable to find the definitive answer to how a soft skill task of management or leadership is performed. The analyst would only continue to refine procedures. This would not help the training developer, and would not aid the critical task selection board. On the other hand, the training developer can approach the process from a different perspective. The developer's objective is to come up with instructional models. Soft skills involve transfer abilities, and the case study method is an appropriate way to teach these skills. To do such training development and teaching requires in-depth knowledge by the instructor. It therefore makes practical sense to give that person a greater range of responsibility.

Just shifting a part of the traditional analysis function to the training developer does not solve the difficult problem of teaching soft skills. Perhaps no system will ever do that. Such teaching must rely on expert and talented instructors, imagination, flexibility and interchange among students and teachers. The shift does prevent wasted energy, however. The analysis phase for soft skills does not need to extend so far as for hard skills, and the training developer needs to become involved earlier. The task selection process can be done at the duty level, and for soft skills this is the proper place.

THE ANALYSIS OF "SOFT SKILLS": AN IMPLEMENTATION STRATEGY

CPT Robert R. Begland

HQ TRADOC
Training Developments Institute
Fort Monroe, VA 23651
AV 680-3608

INTRODUCTION

The analysis of soft skills in the Army and other services has been either ignored or accomplished with whim and fancy. Because of its difficulty and vagueness little attention has been paid to how a soft skill analysis should be conducted. The procedures of how to conduct a soft skill analysis are not to be found in the research. The previous two papers represent approaches to conducting soft skill analyses at both the Military Police School and at the Judge Advocate General's School. They represent locally developed initiatives that attempted to resolve a truly difficult analysis problem. The paper presented by LTC Walton, describing why the Training Developments Institute is involved in this analysis effort, attempted to indicate the extent of the problem and the specific purpose for a soft skill analysis model. The last paper to be included in this program will present a draft soft skill analysis model that has been developed and is presently being researched by the Occupational Research & Analysis Division. The purpose of this paper is to describe the procedures and strategies developed by the Training Developments Institute to achieve institutionalization of that model.

PURPOSE OF PROJECT

This soft skill analysis project had as a goal the development of a procedural model that would fill an identified analysis gap. The existing procedures in the Inter Service Model described steps to be followed when doing a job and task analysis. But the analyst quickly finds out that these steps do notadequately describe the activities required in a complete analysis. When the analyst begins to look at some of the "soft skills" that he encounters during the job analysis it quickly becomes evident that the existing procedures are inadequate. The model under development is designed to provide to the Army service schools a procedure that will truly improve the practice of job and task analysis to an extent that the school analyst will be able to conduct a comprehensive analysis and produce the appropriate analytic base about which training development efforts can be supported. Recognizing the lack of a procedural model for the conduct of soft skill analysis it is hoped that this research effort will be an advancement of the theory and contribute to the training development efforts of the Army in the future. Yet we in the Occupational Research & Analysis Division of the Training Developments Institute can not be satisfied if we have produced a procedural model that is not utilized by the schools. A quick review of the major educational innovations of the past 25 to 50 years points out that the majority of these "good ideas" were never institutionalized. Success in this soft skill project can only be identified through the implementation of the model and an improvement of the training system and products which result from this analysis base. Institutionalization of the model is an essential component to advancing the state of the art.

POTENTIAL PROBLEMS

Limited Research

In most training development areas, the educational researcher is able to turn to a substantive research base.In the area of soft skill analysis one discovers that a lot of people have identified the problem but few people have ever decided to do anything about it. An extensive review of the research indicated that there has been little work done in the area of soft skill analysis. Aside from a conference held in 1972 at Fort Bliss, Texas by the Army's Continental Army Command, there has been little published in this area.

Inadequacy of Existing Procedures

The present reference for all interservice training development efforts is the "Interservice Procedures for Instructional Systems Development" developed by Florida State University under contract with The Combat Army Training Board. This interservice model was intended to produce a generic systems model for instructional development. The Army implemented these procedures as its TRADOC pamphlet for the accomplishment of training development. Although this model certainly provides the framework and

Best Available Copy

structure for the accomplishment of training development, its application in each of the various phases leaves the analyst and designer with insufficient information to design training. Specifically in the area of Job and Task Analysis, the analysis phase of the IPISD materials were deemed to be inadequate for traditional job and task analysis and did not address at all the area of soft skill analysis.

Quick Fix Philosophy

Ideally a researcher is always provided sufficient time to accomplish the research effort. Realistically in the services and many other educational areas it always seems that the training development effort had to have been accomplished yesterday and the analysis effort should have been done two years ago. Recognizing that this is the rule rather than the exception it was obvious that the schools in the Army's training system have been clamoring for years for guidance on how to do soft skill analysis. In an environment where the sentiment is always, "I should have had it done yesterday", it is essential that a quick fix approach to soft skill analysis not be chosen as the preferred strategy. Quite the contrary, a systematic approach to the development, validation and evaluation of a procedural model should be accomplished over a period of time that would allow for the development of a model that would work, vis-a-vis, a model that was developed over a two week period.

Experts vs Expertise

One finds that if you canvas the service schools and ask for individuals who are experts in the area of soft skill analysis, there are many people who identify their professional competencies and would label themselves as expert in the field of soft skill analysis. Yet a more detailed discussion with such individuals points out that the actual expertise in the area of soft skill analysis is minimal.

Bureaucracy

The Army's individual training organization is built upon approximately 23 Army service schools that conduct training throughout the United States. These independent autonomous activities are supported by the Army's Training and Doctrine Command that is located at Fort Monroe, VA. The guidance and procedures that are put forth by TRADOC are then acted upon independently by each of the Army's service schools. This bureaucratic structure does not necessarily lend itself to achieving institutionalization of any type of a model. Recognizing the autonomy of each of the Army service schools any procedural model that is presented to these institutions will be viewed with a jaundice eye. The typical response of "not invented here" may be a potential problem for achieving institutionalization of said model.

ISD:Protagonist/Antagonist

The Army's service schools have adopted a systems approach in their training development efforts. When one says ISD in the Army it is to some people a religious term. Inherent in any spiritual or religious activities there are certain traditions or rituals for doing things. The analysis phase of ISD has these ritualistic activities and there are zealots in the schools who are inflexible in modifying their behavior to learn a new procedure. Thus any procedural model which is a modification or extension of the existing ISD procedure will certainly be met with hesitance if not violent rejection by these ISD protagonists. Within the Army's service schools there are those ISD antagonists who point out that the problem with the Army's present training approach is directly attributable to all of those "educational technologists" who have ruined Army training by their systems approach model. These individuals (the ISD antagonists) will perceive any soft skill analysis model as being associated with the systems approach and will thus reject the model because of their resentment of the ISD philosophy. The essence being you will be damned if you are and damned if you ain't.

The Almighty Task

One of the tenets in a systems approach to training is that the training developer specify a priori precisely the learned capability the person must possess as a result of training. In Magerian terms this simply means that the training developer will specify for the student the task, condition, and standard for performance. This training objective is then provided to the individual in advance of instruction to inform him of what it is he has to be able to learn or do as a result of training. This preoccupation with task specificity has in fact contributed substantially to the analysis and design problems because the task has now taken on its own identity and has been perceived as an end in itself, as opposed to a means to achieving the trained soldier. The soft skill analysis process is a modification and extension of the degree of specificity found in the traditional task, condition, standard format.

IMPLEMENTATION CONSIDERATIONS

Acknowledging that a substantial percentage of the educational innovations of the past never achieved institutionalization, a review of their implementation strategies points out a series of general rules to be aware of when designing a dissemination and diffusion strategy. These implimentation considerations are provided as a strawman about which to develop the specific strategies for achieving institutionalization.

You cannot disseminate an idea.

People are resistant to change.

Institutions have parochial interest.

Innovations always require more effort than the present system.

People must know about a product before they will adopt it and perceive that it will meet some personal need.

Premature information about a project or production often does more harm than good.

Awareness activities should describe what has been done, not what is going to be done.

The effectiveness of awareness activities usually depends upon generating adequate multiplier effect.

Demonstration by the developer is the best means for making people aware of an innovation and convinced of its personal utility. Personal contact with a familiar person who is an advocate is next best.

Awareness materials and activities should clearly delineate the advantages that will accrue to the adopter.

Dissemination of innovations now depends heavily upon convincing and training large numbers of diverse people.

With these implementation considerations in mind, the strategies for gaining institutionalization become a little clearer. Yet there are several specific characteristics of a given system that make that system resistant to change.

RESISTANCE TO CHANGE

The change agent working within an organization must recognize that there are several characteristics of people and organizations that are historically present and which will preclude the achievement of a specified change in behavior unless these characteristics are accommodated.

The structure of the organization

Change apparatus lacking

Lack of incentives for change

Commitment to status quo

Ignorance

Misinformation

Fear of unknown

Previous efforts

Ex innovators

These barriers to change will be present in most organizations and will constitute one of the major areas which the change agent must address in achieving institutionalization of any procedural model.

DISSEMINATION/DIFFUSION MODELS

Most educational innovations have been produced through a major dedication of research dollars to the development of a set of instructional materials with the accompanying teacher training programs. Once this set of packaged materials are produced they are then normally provided to the establishment and a different group of personnel begin the actual dissemination and diffusion of these packages, hoping that the system is ready to accept these innovative concepts. This traditional approach to the design and dissemination of educational innovations could be labeled as Phase A and Phase B where Phase A indicates the preliminary design effort and Phase B is the subsequent dissemination and diffusion effort which occurs after the design activity, separate from and accomplished by a different group of people. This approach has been a primary cause for the lack of institutionalization of most educational innovations. A more appropriate design, development, dissemination and diffusion strategy is portrayed on a model wherein the development activities and the dissemination and diffusion activities occur simultaneously. This model proposes that the development efforts for an educational innovation decrease over time yet the dissemination and diffusion efforts increase over time. The significance of this model is derived from the fact that there are dissemination and diffusion efforts which begin simultaneous to the development efforts.

STRATEGIES

School Involvement

Any change agent in an organization realizes that it is essential to involve critical/key players from the beginning for gaining acceptance of a model or procedure. From the beginning of this project the TRADOC has involved the schools in the soft skill analysis effort. Beginning with problem identification, solution identification, where specific schools were provided an opportunity to collaborate in the development of the research design for this model. Early on all of the Army service schools provided comments concerning their perceptions of what they though the soft skill analysis area was, and what were the major considerations in the conduct of soft skill analysis. If the individual schools do not feel that they have been kept informed about what has been happening, then in fact their involvement and the appropriate information flow will prevent the institutionalization of this soft skill analysis model.

Information Flow

In order to ascertain what was the existing state of the art in the area of soft skill analysis, in July of 1979 TDI sponsored a symposium that brought together a variety of individuals with extensive background in the area of training development and analysis with the extensive participation from Army schools. The purpose of this symposium was to create a think tank in which a potential model could be drafted and which would then subsequently be developed by various consultants in collaboration with the Training Developments Institute. This symposium and its participants and the model were discussed by LTC Walton.

The Chiefs of Analysis Seminar has been an activity in which the TRADOC HQ has communicated to the chief of analysis within each school (who is responsible for the supervision of all analysis efforts) to convey to these individuals, models, procedures, and job aids that facilitate the accomplishment of their day-to-day task. This twice a year seminar has been used as a substantial information dissemination activity within which detailed information has been put out.

Through personal newsletters from the Chief of Occupational Research and Analysis to the key players in each of the Army service schools we have periodically informed the individuals where we were at on the soft skill analysis project, thereby controlling the information flow and providing only the required information. In terms of presentations external to the normal service school contacts, the Training Developments Institute has scheduled presentations at several major conferences to inform agencies external to the service schools what the organization was doing in the area of soft skill analysis so as to facilitate subsequent acceptance by other agencies, and to provide input to these other organizations what the Army has been doing in this area.

Within each service school there are senior officers who are responsible for the total training development effort, the Training Developments Institute scheduled briefings at major command conferences at which the soft skill analysis project has been described in detail so as to provide these senior managers a concept of what has been done, why it is being done, and where the program will go. Recognizing that the soft skill analysis effort has gone from a preliminaiy concept developed in June, 1979 through a research design model with potential implementation on a limited basis in March 1980 there are several evaluation programs and disseminations which have been scheduled for that time frame. The training materials and the workshop materials that are required for development to support such a major analysis effort are presently under development and will be provided to the field on a draft basis for formative evaluation during the January-February time frame and an external evaluation effort is tenatively projected for the summer and fall of 1980.

CONCLUSION

This type of a dissemination diffusion strategy has been designed to anticipate the projected problems in trying to institutionalize a soft skill analysis procedural model in the Army's 23 service schools. Recognizing that the ultimate goal is the improvement of individual training, it is obvious that the best designed training development model will fail if the service schools choose to reject or ignore the model. For this reason a substantial dissemination diffusion plan has been developed to attempt to accommodate any and all potential problems and to thereby facilitate the ultimate institutionalization of a soft skill analysis model that can be used within the schools and will in fact produce sufficient analysis data that is beneficial to the total training effort.

TRANSFER TASK ANALYSIS

Paper presented at the Military Testing Association Conference

San Diego, California October 19, 1979

M. David Merrill

Professor of Education, University of Southern California

In recent years Army training has become more performance oriented. Actual job performances are described in the form of task statements and soldiers are taught to perform these tasks. With the increased complexity of modern equipment and procedures, however, two problems have become apparent. First, the number of relevant tasks continues to increase to a point where learning all of the necessary tasks is almost beyond a given soldiers available training time. Second, many jobs require performance which is difficult to characterize as procedural.

The purpose of this paper is to describe an analysis procedure which is appropriate when the job to be done requires more than the execution of a specific procedure. This analysis procedure is called transfer task analysis.

What is transfer?

The best way to understand what is meant by a transfer task is to compare it with a procedural task. A procedural task is one (1) for which there is a fixed routine for doing the task, (2) for which the desired result will occur if the routine is followed, and (3) for which the desired result will not occur if the routine is not followed. On the other hand, a transfer task is one (1) for which there are at least two approaches for doing the activity, (2) for which following a set procedure that yields the desired results in some situations may not yield thedesired results in other situations or (3) for which two or more sets of procedures may yield the desired results in a single situation.

Unlike procedural tasks, transfer tasks are rarely taught as simple routines. They are taught as rules, laws, or principles which must then be used to derive a unique procedure for a given situation. The occassion for the application of a transfer task is unlikely to be the same for different performers at different times. These differences are the reasons for a transfer rather than a procedural approach to performance.

Transfer tasks versus "soft skills"

There has been considerable dialogue about instruction and analysis of so called "soft skills". Transfer tasks represent only one type of soft skill. Problems posed by other soft skills require techniques not described in this paper. In addition to transfer tasks at least three other types of soft skills have been identified: motivation problems, selection problems, and assessment problems.

There are many problems which cannot be solved by training. For example, if employees continually come to work late a training program in, "How to come to work on time", is unlikely to solve the problem. It isn't lack of knowledge which prevents them from being on time but some problem associated with their motivation. Analysis may be required to determine the cause of the motivation problem and to derive a potential solution, however, this is not a transfer task and the solution is not transfer task analysis.

There are some problems for which no amount of training will solve the problem or where the amount of training required to modify the characteristic is beyond the reasonable time limits for the training system. These personality traits require years to acquire and, while subject to change through experience, are not likely to be changed in the short period of time typically available for training. When such characterists are critical for a particular job environment it is best to test applicants and select those who already have the desired characteristics rather than trying to train candidates who do not have these characteristics.

A third category of problems involves those tasks which are difficult or dangerous to access. The solution to these problems is with the measurement techniques involved rather than in the analysis for training. The solution to such problems often lies in the area of simulation or gaming. These types of procedures are of a different type from those considered in this paper.

How to decide if your training problem involves a transfer task

Figure 1 illustrates a decision process for determining whether or not a given instructional task involves transfer and thus requires transfer analysis. The first step in any analysis procedure is to determine the nature of the task. This is usually accomplished by a performance description. A performance description indicates what the person does or accomplishes as a result of the performance.

As indicated in figure 1 the first decision for the analyst is to determine whether or not the performance described can be acquired by training. If adequate performance is a matter of appropriate motivation or the result of some personality characteristic then some other analysis techniques should be applied.

Having determined that the performance is subject to modification by training the next decision for the analyst is to determine whether there are variations in the way the task can, or should, be performed in order to produce an acceptable outcome. If there is only one path through an acceptable procedure then it is a unitary task and conventional task analysis whould be appropriate. If there are several acceptable procedures to accomplish the outcome or if there are several paths through a given complex procedure then it may be a transfer task.

Having determined that variations exist the analyst may now observe that the task in question has only one procedure, but that this one procedure has two or more paths that can be followed to produce an acceptable outcome. On the other hand the analyst may observe that the task has two or more procedures that can be used to produce an acceptable outcome and that any of these procedures may have more than one path. If the tasl has only one procedure with multiple paths then it is a multiple procedural task. Conventional task analysis should be appropriate.

On the other hand if the task has two or more procedures then it is a potential transfer task. However, it may still be inadvisable to use transfer task analysis. To make this decision, there are two questions the analyst should ask: (a) Do different situations require different procedures? (b) Can different procedures be used in the same situation? With respect to the first question, more than one procedure may be required because the same procedure does not produce acceptable outcomes in all situations in which the task is performed. This is called situation variation. With respect to the second question more than one procedure may be required for doing a task because different individuals may prefer to use different procedures in a single situation. This is called performer variation. Figure 2 illustrates the relationship between performer variation and situation variation.

As illustrated in figure 2 performer variation and situation variation represent continuous dimensions. To decide whether or not to use transfer task analysis the analyst must ask, "How many different procedures can be used?" If only a small number of procedures can be used it may still be best to analyze each of the individual procedures using conventional task analysis. Each individual procedure would be taught.

However, if there are many procedures that can be used in is more appropriate to use transfer analysis which will determine the basic principles or theoretical knowledge from which all of these procedures can be generated. This underlying knowledge becomes the focus of instruction rather than each of the individual procedures.

A performance description is usually stated in terms of observable activity. However in a transfer task the observable activity may not reflect the mental activity necessary to derive the necessary observable performance. For example, consider the task " assigns work loads". The actual assignment is a straightforward matter. The difficult part of the task is to figure out what jobs to assign to which people, how to be equitable, how to insure work efficiency, etc. The mental decisions required vary with the personnel available, with the situations, and with the individual making the assignments. Each new work assignment opportunity requires the individual to derive specific procedures which will work in that situation at that time. This derivation of appropriate procedures for the actual assignment is difficult to observe.

Transfer task analysis focuses on analyzing the basic knowledge from which all of the necessary specific procedures can be generated. This basic knowledge is referred to as a theoretical model. If an acceptable effective theoretical model is not known by anyone anywhere, then it cannot be identified by transfer task analysis nor can it be taught. In this event the appropriate action is to recommend a research project to develop an acceptable theoretical model.

If an acceptable effective model is known then the analyst should use the transfer analysis process described below. For any task there is a known-to-unknown continuum for the underlying theoretical models. Something is usually known but not everything. Hence the analyst must judge whether or not the existing model(s) are sufficient for the task at hand or whether more work is needed to discover appropriate models. In many cases, because the task must be performed,the analyst will be forced to accept whatever models exist and do the best to make them work because time and resources will not allow him/her to wait until adequate models can be discovered and tested.

How to describe a transfer task

Figure 3 identifies the components of a transfer task. As has been described in the previous paragraphs a transfer task involves two stages: (1) The use of an underlying model to derive a procedure unique to a given situation or person; and (2) the application of this procedure to perform the task. When the procedure can be directly taught then the first phase can be ignored, but when a new procedure needs to be derived for each unique situation then the analyst must be concerned with specifying the underlying model and teaching the student to use the model to derive procedures as well as teaching the student to use the procedures once they have been derived.

The identification of the model underlying a transfer task can be accomplished in at least two ways. In the bottom-up approach the procedure for a giver person and/or situation should be carefully described in a way similar to that used for conventional task analysis. Then a second procedure should be described for a second situation or person. This second procedure should differ from the first while still accomplishing the objective of the performance. The analyst then examines the two procedures carefully in an attempt to identify the commonalities and the cause and effect relationships which underlie these common elements. It may be necessary to derive a third procedure for this analysis process. After the knowledges, principles, etc., which form the underlying model have been specified another procedure for yet another situation and/or person should be derived and tested against the model to be sure the model is sufficient to generate this new procedure.

The top-down approach to deriving models is the academic approach. Start by trying to identify the model from known sources such as texts, manuals, experts, etc. Once the model has been specified use it to try to derive a procedure for a given situation or person. Once it has been demonstrated to be adequate in one situation use it to derive a second procedure. This is a conceptual test of the model. The problem with the top-down approach isthat the model specified often contains nice-to-know information which is not necessary for the student to be able to derive necessary procedures. There is a tendency to teach all there is to know rather than that which is necessary to know.

In practice a combination of the bottom-up and top-down approaches should probably be used. In this way the model derived is likely to be complete without containing unnecessary information.

It should be notэd that in instruction the student must learn more than the model. The instruction should teach two levels of skills. First the student should be taught to use a procedure which is appropriate to a given situation/person combination. This ensures that s/he can use at least one example of the procedure. Then the student should be taught the model which underlies the procedure which s/he has learned to use. The student must then be taught to use the model to derive a second procedure appropriate for a different situation/ person combination. The student should then be given practice in using the procedure that s/he has derived in this new situation. This practice at deriving procedures and applying them should be repeated until the student is able to derive new procedures for a variety of situations. The number of such repetitions will vary depending on the complexity of the model and the resulting procedures.

Some problems yet to be solved

This paper is very general in terms of the derivation of models and the form that these models should take. A major problem for the specification of transfer task analysis is how to identify the underlying competencies which should be taught. What do adequate models look like? How does one represent the interrelationships involved?

A second, but related, problem is how to structure the principles which comprise a model so that memory is maximized, transfer to new situations is maximized, etc.

A third problem is how to teach a student when a given combination is appropriate and when a given combination is not appropriate. How does one make the judgment process used to derive procedures as systematic and reliable as possible without resorting to teaching each of the possible procedures.

A fourth problem is how to deal with "quirks". No matter how good a given model ma be there are always situations where some unique element in the situation must be taken into account in order for a given procedure to work. How should such quirks be identified and how should they be taught to the student?

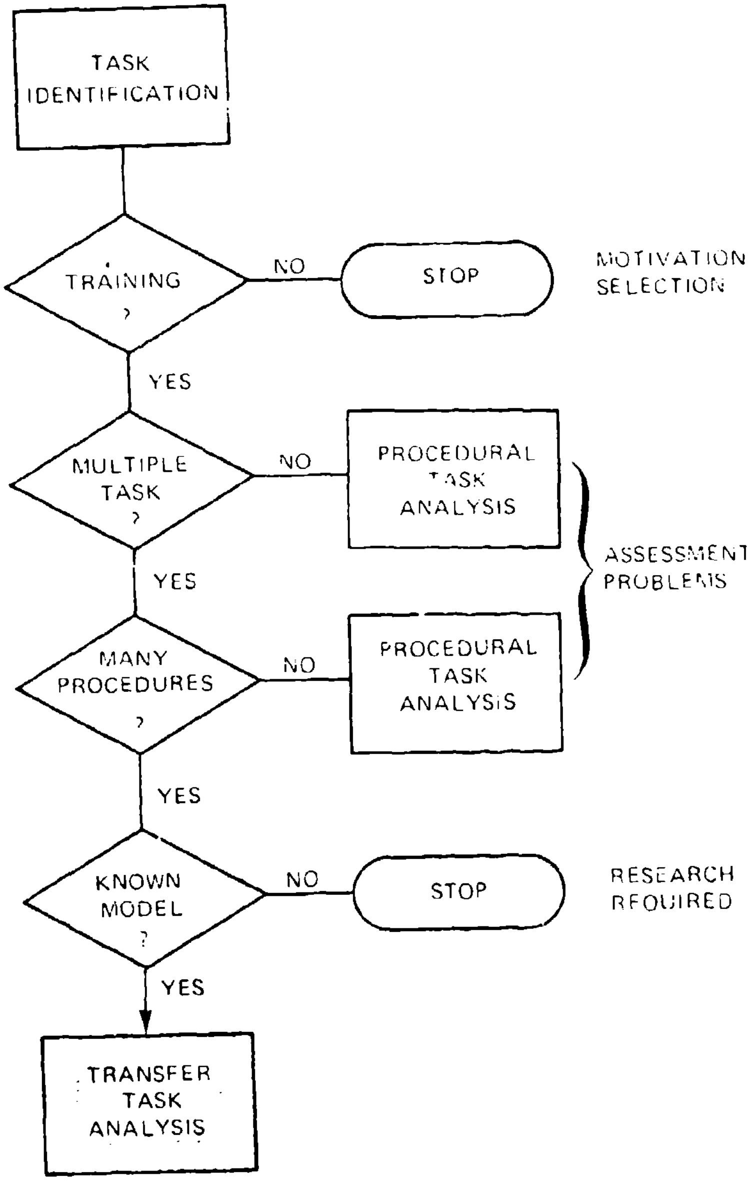
There are no doubt other problems. Much work still needs to be done before the transfer task procedures or the transfer task model form which the procedures can be derived is sufficient to enable analysts to adequately handle transfer tasks. Nevertheless, we feel that the ideas presented in this paper take the first step toward such analysis procedures.

Conclusion

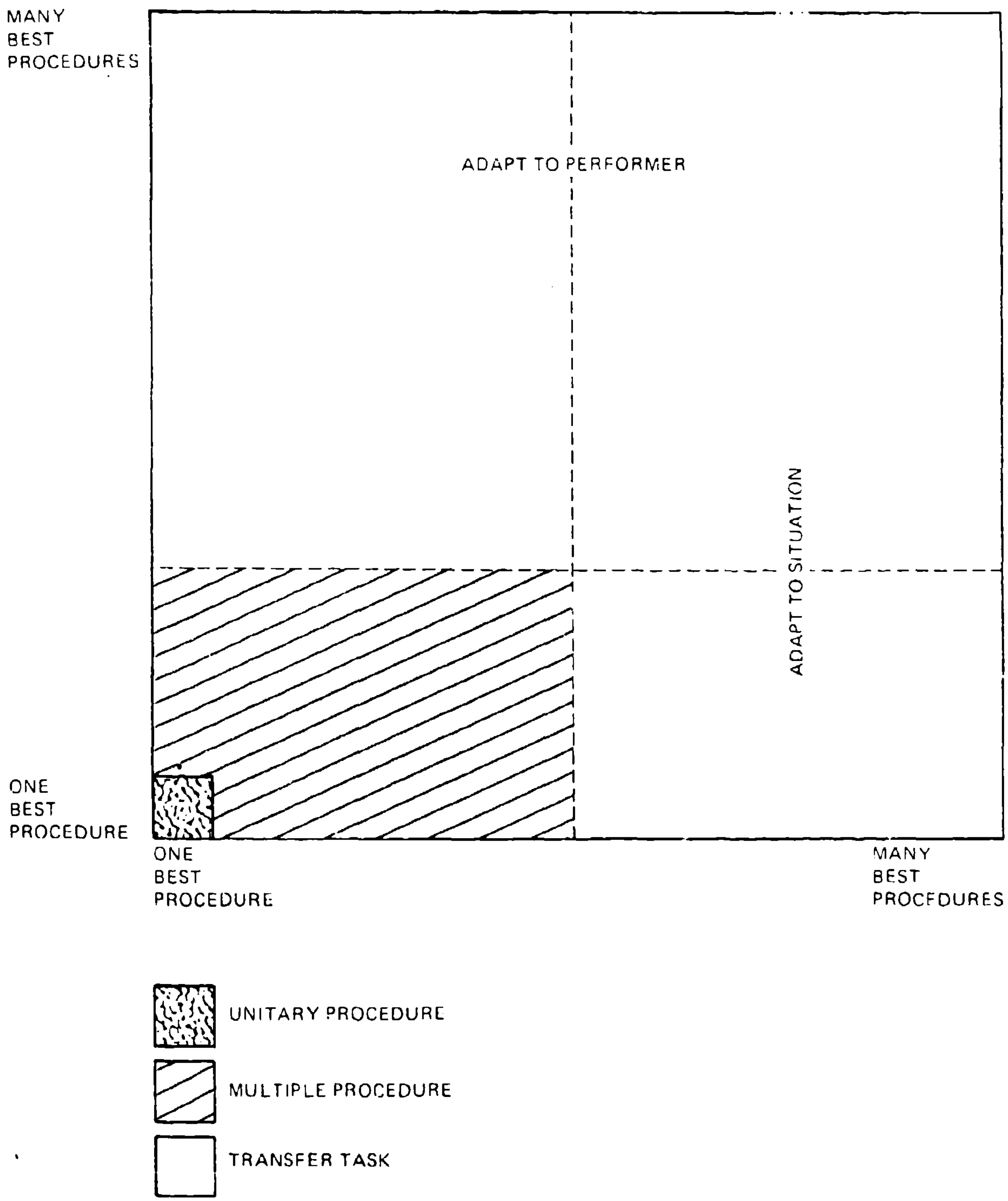
One of the characteristics which has separated "educatior" from "training" has been the claim of educators that they are teaching theory which can be applied in a variety of situations whereas trainers have usually claimed to be teaching specific jobs. This paper is an attempt to indicate that neither extreme will get the job done. An attempt to train specific procedures bogs down in our complex world because of the large variety of specific procedures needed when situationa and personnel change. On the other hand an attempt to teach only theory leaving the derivation of specific procedures to the student in many situations is inefficient and costly. The fine line for instructional analysts is to determine where training of specific procedures is cost effective and where a more educational approach which teaches underlying theory is actually more cost effective. This paper is an attempt to help instructional analysts with this decision process.

Footnote:

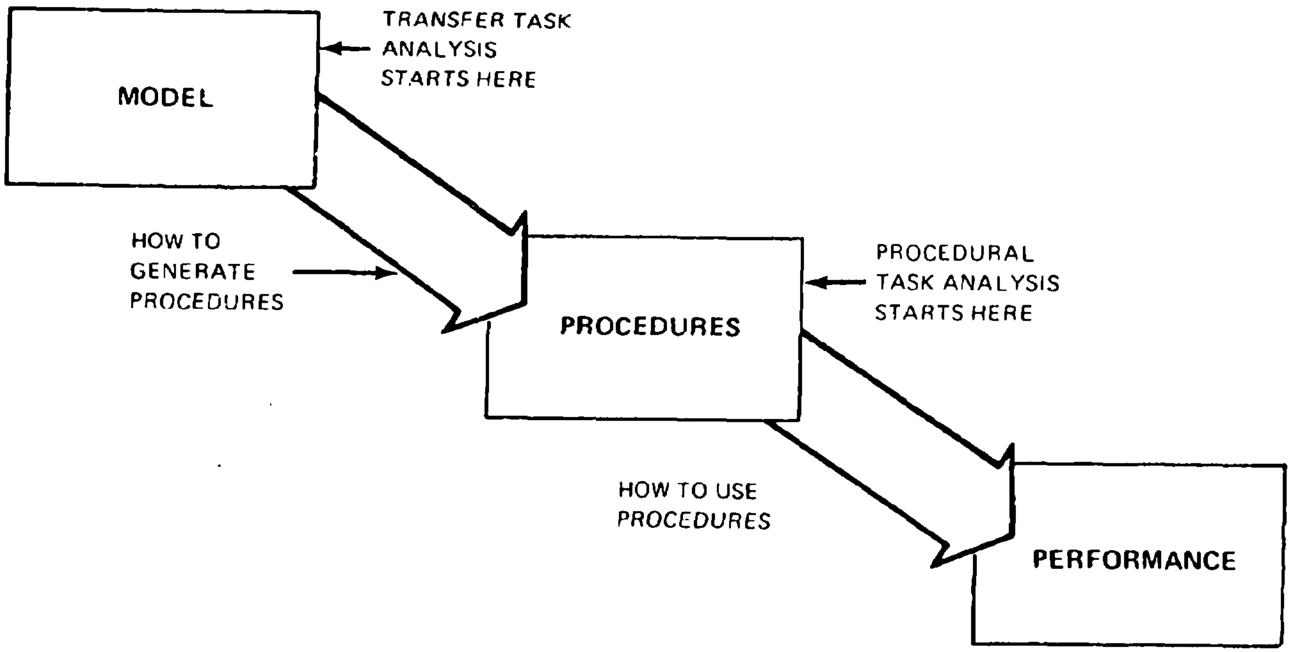
Some of the material for this paper was adapted from a draft of the chapter "Job and Task Analysis Handbook -- Chapter 9 -- Transfer Task Analysis" which was prepared by Charles M. Reigeluth, Ivan Horabin, and Robert Branson.



SELECTING APPROPRIATE TASK ANALYSIS



PERFORMER VARIATION AND SITUATION VARIATION


MODEL
TRANSFER TASK ANALYSIS STARTS HERE
HOW TO GENERATE PROCEDURES
PROCEDURES
PROCEDURAL TASK ANALYSIS STARTS HERE
HOW TO USE PROCEDURES
PERFORMANCE

COMPONENTS OF TRANSFER TASK

OBESITY AS A HEALTH RISK FACTOR

Owen Maller, Ph.D.
Behavioral Science Division/F.S.L.
U.S. Army Research and Development Command, Natick, MA

Alan H. Taylor, Ph.D.
Veterans Administration Outpatient Clinic
Boston, MA

Barbara Edelman-Lewis, M.S.
Behavioral Science Division/F.S.L.
U.S. Army Research and Development Command, Natick, MA

U.S. Army Research and Development Command
Natick, MA 01760

The incidence of obesity in the American population is rising dramatically. National efforts to curtail this increase are being promoted by various federal health agencies and private professional medical societies such as American Heart Association, American Dietetic Association, American Medical Association etc. Despite these efforts, obesity is still on the increase. Current statistics estimate that 30-50% of the American population is obese (1). Unfortunately, no systematic epidemiological survey has been conducted by the Department of Defense or its affiliate branches to assess the prevalence of obesity in the military population. Most available evidence regarding the occurrence of obesity in the military is anedoctal, deduced from current military regulations on physical fitness and weight standards. Despite evidence, it is assumed that the incidence of obesity in the military sector parallels the incidence in the civilian sector. Attempts to relate incidence to age, sex, and duty assignments in the military cannot currently be delineated at this time. It seems reasonable to assume however, that the incidence of obesity is higher in more sedentary duty assignments.

Complications associated with obesity are serious physical disabilities, increased likelihood of contracting degenerative diseases resulting in escalating health care costs, absenteeism and impaired work efficiency.(2) Excessive weight stresses the cardiovascular, respiratory and renal systems. (3) Among the most common complications of obesity are respiratory disorders. Such disorders usually produce a lower exercise tolerance, and a greater difficulty in breathing, the result of decreased vital capacity. A higher frequency of respiratory infections has also been observed. Obesity also increases the likelihood of orthopedic problems thus limiting motor activities. (4)

The interpretations and conclusions presented are those of the authors and do not reflect the official policy of the Department of Defense.

Attempts to understand the relationship of body composition to work efficiency have been limited and unsystematic. It is important to know an individual's or groups' capacity for work performance, particularly when selecting and assigning personnel for special tasks or duties. This paper presents a brief review of the meager research on the relationship of obesity and work performance and its implications for the selection and assignment of personnel.

Young (5) compared the physiological responses of overweight and physically fit adults at rest, during moderately intensive work and during extremely intensive work. The following table depicts difference between the two groups. For relatively intensive work, the physically unfit or obese, show a shorter duration of effort; their maximal oxygen uptake is less despite an increase in pulmonary ventilation. Heart rate is higher, indicating a loss of cardiorespiratory fitness; blood pressure and heart rate slowly decline following work. The oxygen debt and blood lactate levels are reduced suggesting some loss in anaerobic metabolic capacity.

Under moderate work conditions, the physically less-fit must expend greater energy to accomplish the same level of work output. Consequently, there are increases in oxygen consumption, respiration and heart rate. The cardiac stroke volume of the physically unfit is reduced, thereby increasing anaerobic metabolism with a concommitant rise in blood lactate production. The net effect of these physiological changes is the reduction in work capacity.

TABLE (Young)

EFFECT OF PHYSICAL FITNESS IN MAN ON
PHYSIOLOGIC RESPONSES TO WORK

Level of Activity	Index	Relative Change in Unfit Subjects
A. At rest	1. Pulse rate	+
B. Easy work that can be sustained in the steady state	1. O_2 consumption	+
	2. CO_2 production	+
	3. RQ	+
	4. Ventilation	+
	5. Respiratory rate	+
	6. O_2 pulse	-
	7. Ventilatory efficiency	-
	8. Pulse rate	+
	9. Pulse rate deceleration after work	-
	10. Systolic pressure	+
	11. Rate of decline of systolic pressure after work	-
	12. Cardiac output - stroke volume	-
	13. Mechanical efficiency	-
	14. Blood lactate during work	+

Level of Activity	Index	Relative Change in Unfit Subjects*
C. Exhausting work that cannot be sustained in the steady state	1. Duration	-
	2. Maximal O_2 uptake	-
	3. Maximal CO_2 production	-
	4. R.Q.	+
	5. Maximal ventilation	+
	6. Ventilatory efficiency	-
	7. O_2 pulse	-
	8. Maximal pulse rate	+
	9. Pulse rate deceleration after work	-
	10. Systolic pressure	-
	11. Rate of decline of blood pressure after work	-
	12. Blood lactate at end of work	-
	13. Blood sugar at end of work	-
	14. O_2 debt	-

* + or - refers to relative increase or decrease as compared to a physically fit subject.

In a study conducted by Prentiss (6) 21 obese college women were placed on an exercise regimen using a bicycle-type exerciser. The exercise schedule called for a 15 minute ride with three evenly spaced 30 second work periods during the ride. As individual performance progressed (improved) the number of work periods were increased. As weight was lost, endurance capacity was significantly increased. Significant decreases in heart rate were also noted despite the sharp increase in total work output.

Brady et al. (7) examined the relationship between body fat and gross motor proficiency. Ten tests of gross motor performance were administered (8). These included: extent flexibility test-ability to stretch or rotate the spine as far as possible; dynamic flexibility test-ability to make repeated rapid flexing movements in which the resilience of the trunk and back muscles is critical; explosive strength factor-ability to mobilize energy for bursts of effort; static strength factor-the maximum force that can be exerted against an external object (e.g. lifting weights, pulling, pushing); trunk strength factor-more limited to specific trunk muscles; gross body equilibrium-ability to maintain equilibrium, despite force pulling one off balance; and stamina factor-the capacity to continue maximum effort involving the cardiovascular system. Body weight was not significantly related to performance on measures of balance, explosive strength, dynamic flexibility, extent flexibility, trunk strength and static strength. The data demonstrated, however, that the greater the level of body fat, the greater the impairment in dynamic strength, gross body coordination and stamina. These three measures suggest that obese individuals are at a work disadvantage, even before physiological reserves are called into play during stressful periods.

In a study conducted on the relationships of physical fitness, work load and mental performance by Sjoberg et al. (9) two groups of 24 trained and 24 less trained male students participated in an experiment involving two mental tasks, performed under five different work-load conditions: sitting still, pedaling on a bicycle without a load, pedaling with a 25% load imposed, pedealing with a 50% load imposed, pedaling with a 75% load imposed. The load was based on individual established maximal working capacity. Task 1 included demands on continuous concentration and the switching of attention requiring short-term memory. Task 2 involved paired associate learning with recall following short and long retention delays. The physically-fit group performed significantly better on Task 1 across all levels and work conditions. In Task 2, no differences were found. The investigators concluded that well-trained individuals perform better than physically unfit candidates on tasks requiring an intense and sustained concentration. Task 2 demanded little concentration and less mental effort hence the lack of differences in performance.

The above studies suggest performance is affected by weight status. High levels of body fat result in an inefficient utilization of energy, loss in cardio-pulmonary function which results in decreased endurance, stamina, attention-span and information processing. The relationships among these variables needs to be further studied.

Military preparedness requires that a soldier be in a state of combat readiness, i.e. physically fit to perform his/her tasks during sustained periods of stress. Current war scenarios include the performance of complex mental tasks, under conditions of limited physical activity characterised by prolonged stress and isolation. Based upon the preliminary studies, one deduces that the physically-unfit, obese soldier will not optimally perform.

The paper presented by Dr. Johnson represents a starting point for further research efforts. Attempts to develop weight standards based upon body composition measures rather than the traditional but inadequate height and weight standards are disputable as valid indices of body fat levels. This work is a first step in determining the prevalence of obesity in the military. After such standards are established, studies will be required to assess quantitatively the effects of obesity on the performance of military duties, under combat simulation, as well as in garrison. Along these lines are two studies being presented today.

Dr. Kowal's paper on, "Body Composition and Its Relationship to Injuries in Female Trainees," uncovers a number of physiological and prior-fitness measures correlated with injury susceptibility. His observations offer an opportunity for developing remedial programs to minimize the orthopedic and medical consequences associated with military physical conditioning. In addition, his research demonstrates that valid and reliable work performance studies can be conducted under actual field conditions.

In Dr. Hodgdon's study on the effect of carbohydrate loading on endurance capacity, dietary intake as it relates to work performance is examined. Their research shows that selective dietary manipulations combined with specific training activities can yield significant improvements in performance even under periods of prolonged metabolic stress. This study underscores the role of dietary intake and specific nutrients to performance. Furthermore, it emphasizes the need for research on the relationship between specific nutrients requirements and performance.

As we begin to understand the relationships between body composition and performance, we will be able to recommend specific nutrient requirements that will lead to improved performance.

In conclusion, adiposity is important not only because of its well-established association with health-risk factors but physical performance. The effects of obesity on performance are a long-term consequence of certain dietary intake patterns. As the results of Hodgdon's Study have suggested, there can be short-term effects of specific nutrient intakes on performance. Both the short- and long-term effects of dietary practices on performance need to be understood for more effective selection and training of personnel for combat readiness.

References

1. Bray, G. A. (ed) Obesity in Perspective, in Fogerty International Center Series on Preventive Medicine. Washington: DHEW Publication (NIH) 75-708, 1975.

2. Armstrong, D. B. 1951 "Obesity and Its Relation to Health and Disease," JAMA, 1951:147.

3. Bray, G. A. The Obese Patient. Philadelphia:Saunders, 1976.

4. Physical Fitness Research Digest, Series 5, No. 2, April 1975.

5. Young, D. R. Physical Performance Fitness and Diet, Springfield, Illinois, Charles C. Thomas, 1977.

6. Prentiss, G. "The Effect of a Progressive Program of Exercycle Exercise on the Cardio-respiratory Endurance and Anthropometric Measurements of Obese College Women," Master Thesis, University of Washington, 1964.

7. Brady, J. I., Knight, D. R. and T. E. Berghage, "Relationship Between Measures of Body Fat and Gross Motor Proficiency," J. Appl. Psychology, 62,2 1977:224.

8. Fleishorman, E A. The Structure and Measurement of Physical Fitness. Englewood Cliffs, N.J. Prentiss-Hall, 1964.

9. Sjoberg, H., Chlscon, M., and S. Dornic. "Phisical Fitness Work Load and Mental Performance," Rep. Dep. Psychol., niver Stockholm, 1975, No. 444.

BODY FAT VERSUS BODY WEIGHT AS A MILITARY STANDARD

Herman L. Johnson & Robert D. Fults
Division of Nutrition Technology
Letterman Army Institute of Research
Presidio of San Francisco, CA 94129

The Army Regulation AR 600-9 (1), combines the Army Physical Fitness and Weight Control Program and includes military appearance as an equal component of the overall program. Obesity is defined as a medical term which indicates an excessive accumulation of adipose tissue manifested by increased body weight and impying excessive caloric intake, a sedentary existence or both. Overweight exists when an individual's body weight exceeds the maximum allowable for his/her height in the Weight Tables for Army Personnel.

The objectives of the Army Physical Fitness Program are to develop soldiers who are physically capable of performing their duties in a combat as well as peacetime environment and to sustain good health and physical fitness through exercise programs. Physical fitness signifies health, skill, endurance, rapid recovery from fatigue, motivation and self-confidence. According to this regulation, physical fitness is part of the individual's professional qualifications and will be considered as such by commanders during evaluations.

The objectives of the Army Weight Control Program are to maintain the weight of all personnel at a level which is best suited for them to perform their duties in a peacetime or combat environment and to present a smart military appearance expected of a combat-ready Army. Commanders are responsible for continuously monitoring all members, both officers and enlisted, to insure that they maintain proper body weight. Improper weight distribution may be grounds fo. a physician's determination of obesity. Excess body fat is seriously detrimental to health, longevity, stamina and military appearance. Members of the Armed Services who are overweight or obese must accept the personal responsibility for weight reduction and control and for physical appearance. The regulation attempts to distinguish between overweight and obesity and states that a member's weight exceeding the maximum for his or her height will not be the sole criterion for a classification as obese and conversely, a member whose weight does not exceed his or her maximum may be obese. The regulation continues to precisely differentiate between overweight and obese as follows: "Evaluation of the body build, muscular development and bone structure may be necessary to differentiate between these conditions. A view of the entire body should be taken, noting the proportions, symetry of the various parts of the body, chest development, abdominal girth, and the condition and tone of the muscles. An overweight member, who is obviously active, of firm musculature, evidently vigorous and healthy and who presents a satisfactory military appearance, should not be classified as obese. Obesity will be determined by a physician at the medical treatment facility." These are all subjective criteria and the ultimate determination will be made by a physician using the same criteria and height and weight. Possibly, he will measure skin-folds using a gratis skin-fold calipers provided by some pharmaceutical company, with the cost closely reflecting its value. Even the physician who has invested 2 or 3 hundred dollars in a good quality calipers probably is not familiar with the exact sites and positioning of the calipers required to obtain reliable results. Experts in body composition use different sites and equations for estimating body fat from skin-folds and each of these scientists develop their techniques by standardizing against at least one other technique that is considered more precise. However, it should be apparent that this determination of obesity is quite subjective. Even requiring that a physician make the determination does not assure any improvement in objectivity or accuracy of the determination.

Physical fitness and weight control for military personnel appears to have three major goals. First, is to assure that the military person is capable of performing his/her duties during both combat and peacetime. Second, is to maintain the health of the person. Third, is that the military person presents a smart soldierly bearing in uniform. The importance of maintaining adequate physical fitness for performance of combat as well as peacetime duties is obvious since the only justification for maintaining military forces is to be prepared for armed conflict. Failure to perform one's duties under such circumstances could result in increased casualties and possibly affect the ultimate outcome of the conflict. The maintenance of good health of the career soldier is important not only for the improved "quality of life", that is important for everyone, but it has implications for the optimal functioning and recuperative capabilities of the combat person and has economic consequences in that the military services have all encompassing medical plans for active duty and career personnel. Therefore, the military is concerned that personnel maintain optimal health in order to reduce health care costs. However, the major emphasis of the physical fitness-weight control programs appears to be appearance and stresses that the wearing of the Army uniform should be a matter of personal pride and satisfaction and that waistlines that stretch the front of the blouse or shirt and "pot-bellies" detract from good military appearance. Recent emphases upon weight control include bar to re-enlistment, unfavorable comments on evaluation forms and strong emphasis of appearance during consideration for promotion. Gross obesity limits performance, is not healthy and detracts from military bearing. Although weight for height is a very easy determination for anyone to make, how much emphasis should be placed upon this standard? Despite the admonitions within the regulation that these tables should not be the sole criterion for establishing a need to reduce, they have become the sole criterion and essentially a sacred criterion, at that.

Body weignt is generally reflective of body fat; however, there are many instances where this is not true. Many athletes engaged in strength-requiring sports may greatly exceed the standards with relatively low levels of fat as has been reported (2,3). The military person with a large body frame and engaged in heavy physical work or recreational activities, may be in this category. To require such a person to lose weight, mainly muscle tissue, would be counterproductive if one's major concern is for performance. The real concern should be for overfat rather than overweight (4) but the methods for estimating body fatness are tedious and generally imprecise, unless utilized by a scientist who has worked with several methods. Since performance is the real concern of the military, the emphasis should be placed upon physical fitness and not on body weight. I would even suggest that body weight standards be altered according to the persons' scoring on their PT tests. A person scoring in the upper 20 percentile might be permitted an extra 10% of weight; the upper 10 percentile, 15% additional weight and the top 5 percentile up to 20% added weight. At the other extreme, those with low PT scores might have their maximum weight allowances correspondingly reduced. This could be another approach to compensation for differences in body frames. It would provide an alternative to weight reduction for the person who exceeded the weight standard and it would be directed to the military goals for performance, health and appearance.

Another aspect of weight reduction especially in conjunction with physical training is energy balance (5,6). Fat _per se_ is equivalent to 9,000 to 9,500 kcal per kilogram and body adipose tissue about 8,000 kcal per kg. Body protein or muscle tissue is equal to 1000 kcal per kg. Therefore a person could be losing 1000 kcal per day, about 1 kg or 2.2 lbs per week of fat and gaining an equal amount of muscle so that his/her body weight is not changing despite tthe caloric deficit. This 1000 kcal deficit may represent 30 to 50% of the caloric needs of the person and to maintain such a diet without observing any weight loss could be rather discouraging. Again, I believe the emphasis should be on performance and health, which is related to fat, rather than weight, _per se_.

It is well recognized that excess fat can reduce endurance performance involving running, which has been attributed to carrying the burden of fat mass. However, a few studies have shown that physical performance capabilities can be increased substantially with relatively smaller changes in body composition (7,8,9). Work of many physiologists has shown awareness of this by attempting to relate work performance to lean body mass rather than total body weight (10).

In some of our studies of nutritional status of military personnel and of nutrient requirements under different stresses, we have measured body composition and work performance (11-15). Body composition measurements, as currently being conducted on West Point Cadets, begin with height and weight and skinfolds (16). It includes body density measurements from water displacement for body fat estimations, 40-potassium counting for body protein or muscle (17) and various tracers for blood volume, total body water and extracellular space (18). Bone minerals can be calculated from certain bone diameters and lengths or from a percentage of lean body mass. From these various measurements, relatively good estimates of the body's muscle, water, fat, mineral and lean body mass can be made. Physical work capacities were determined by using our respiratory gas analyzer (19) to measure total oxygen consumption at rest and through increasing work loads on the treadmill or using the Kofrani-Michaelis respirometers (20) during marching and other military activities or simply recording times required to run 10 to 15 miles. We have recently replaced the respiratory gas analyzer with a respiratory mass spectrophotometer. This mass spectrophotometer, along with the computer, is presently at West Point gathering data on the maximum performance capacity to relate to body composition, and on the relationship of heart rates to energy expenditure. From their relationships we can calculate their daily expenditure from monitors recording total heart beats for 5-day periods. We are simultaneously obtaining total food consumptions using diary interview techniques so that the energy balance and adequacy of nutrient intakes can be determined.

From these data we should be able to: a) determine energy balances of the Cadets during different years at the academy, b) calculate the correlations between body composition and performance capacities, c) sugge acceptable ranges of body composition for Cadets in relationship to physical work capacities, and d) recommend ways to alleviate the problems of excessive weight and weight gains of the Cadets. I have no reason to believe that these data will contradict our earlier studies that oxygen consumption and work capacity are directly related to lean body mass. Therefore, the military standards for weight for height should be replaced by a maximum percent body fat or at least tempered by their physical fitness scores. Since the military's primary concerns for the personnel should be for their performance capacities and health, these can be better assured by relating their physical fitness scores to their weight for height than by maintaining these as separate entities.

1. The Army Physical Fitness and Weight Control Program. AR 600-9, Headquarters Department of the Army. Washington, DC, 30 November, 1976.

2. NOVAK, L.P., R.E. HYATT and J.F. ALEXANDER. Body Composition and Physiologic Function of Athletes. J. Am. Med. Assoc. 205:764, 1968.

3. BEHNKE, A.R. and J. ROYCE. Body Size, Shape and Body Composition of Several Types of Athletes. J. Sport Med. 6:75, 1966.

4. WRIGHT, H.F. and J.H. WILMORE. Estimation of Relative Body Fat and Lean Body Weight in a United States Marine Corps Population. Aerospace Med. 45:301, 1974.

5. GRANDE, F. Nutrition and Energy Balance in Body Composition Studies. In: Techniques for Measuring Body Composition. Eds, J. Brozek and A. Henschel. NAS-NRC, Washington, DC, 1961.

6. GRANDE, F. Energy Balance and Body Composition Changes. Ann. Int. Med. 68:467, 1968.

7. LEEDY, H.E., A.H. ISMAIL, W.V. KESSLER and J.E. CHRISTIAN. Relationships Between Physical Performance Items and Body Composition. Res. Quart. 36:158, 1964.

8. DEMPSEY, J.A. Relationship Between Obesity and Treadmill Performance in Sedentary and Active Young Men. Res. Quart. 35: 288, 1963.

9. HURBREGSTE, W.H., L.H. HARTLEY, L.G. JONES, W.H. DOOLITTLE and J.L. CRIBLEZ. Improvements of Aerobic Work Capacity Following Nonstrenuous Exercise. Arch. Env. Med. 27: 12, 1973.

10. WELCH, B.E., R.P. RIENDEAU, E.E. CRISP and R.S. ISENSTEIN. Relationship of Maximal Oxygen Consumption to Various Components of Body Composition. J. Appl. Physiol. 12:395, 1958.

11. JOHNSON, H.L., H.J. KRZYWICKI, J.E. CANHAM, J.H. SKALA, T.A. DAWS, R.A. NELSON, C.F. CONSOLAZIO and P.P. WARING. Evaluation of Calorie Requirements for Ranger Training at Ft. Benning, Georgia. Report No. 34. Presidio of San Francisco, California: Letterman Army Institute of Research, July, 1976.

12. CONSOLAZIO, C.F., H.L. JOHNSON and H.J. KRZYWICKI. Body Fluids, Body Composition and Metabolic Aspects of High-Altitude Adaptation in Physiological Adaptations. Dessert and Mountain, Academic Press, New York, 1972.

13. CONCOLAZIO, C.F., H.L. JOHNSON, R.A. NELSON, J.G. DRAMISE and J.H SKALA. Protein Metabolism During Intensive Physical Training in the Young Adult. Am. J. Clin. Nutr. 28:29, 1975.

14. CONSOLAZIO, C.F., R.A. NELSON, T.A. DAWS, H.J. KRZYWICKI, H.L. JOHNSON and R.A. BARNHART. Body Weight, Heart Rate and Ventilatory Volume Relationships to Oxygen Uptakes. Am. J. Clin. Nutr. 24: 1180, 1971.

15. KRZYWICKI, H.J., C.F. CONSOLAZIO, H.L. JOHNSON and N.F. WITT. Effects of Exercise and Dietary Protein Levels on Body Composition in Humans. Report No. 62. Presidio of San Francisco, California: Letterman Army Insitute of Research, July, 1978.

16. KRZYWICKI, H.J., G.M. WARD, D.P. RAHMAN, R.A. NELSON and C.F. CONSOLAZIO. A Comparison of Methods for Estimating Human Body Composition. Am. J. Clin. Nutr. 27:1380, 1974.

17. WARD, G.M., H.J. KRZYWICKI, D.P. RAHMAN, R.L. QUASS, R.A. NELSON and C.F. CONSOLAZIO. Relationship of Anthropometric Measurements to Body Fat as Determined by Densitometry, Potassium-40 and Total Body Water. Am. J. Clin. Nutr. 28:162, 1975.

18. NIELSEN, W.C., Jr., H.J. KRZYWICKI, H.L. JOHNSON and C.F. CONSOLAZIO. Use and Evaluation of Gas Chromatography for Determination of Deuteurim in Body Fluids. J. Appl. Physiol. 31:957, 1971.

19. NELSON, R.A., L.O. MATOUSH and C.F. CONSOLAZIO. Development and Application of a Continuous Oxygen Uptake Measurement System. Report No. 318. Denver, Colorado: US Army Medical Research and Nutrition Laboratory, May, 1968.

20. CONSOLAZIO, C.F. Energy Expenditure Studies in Military Populations Using Kofrani-Michaelis Respirometers. Am. J. Clin. Nutr. 24: 1431, 1971.

BODY COMPOSITION AND ITS RELATIONSHIP TO INJURIES IN FEMALE TRAINEES

Dennis M. Kowal, Ph.D.
US Army Research Institute of Environmental Medicine
Natick, MA 01760

Women entering the Army are exposed to considerable physical stress due to the intense physical training program encountered. At the beginning of a basic training cycle a prospective study was initiated to identify exercise related injuries and performance-limiting conditions that resulted and to identify some of the factors that may contribute to their occurrence. Four hundred women recruits, aged 18-29, participated in the study. All had passed an initial physical examination and were without any limiting disabilities. An initial assessment of physical fitness was accomplished to determine the current status of body composition, strength of the major muscle groups, (e.g., legs, trunk, arms and upper torso), aerobic capacity, previous athletic history, self perception of physical fitness, and psychosomatic predisposition. The training and conditioning program consisted of 1hr/day, 5-6 times a week and involved a series of standard warm up calisthenics and stretching exercises followed by a run, beginning at 3/4 mile at a 10 min/mile pace and increasing to two miles at 9-1/2 min/mile by the end of training. Extensive road marches and military training activities were also included. At the end of training a self report injury questionnaire was used to collect injury data. These data were documented with the records from the unit dispensary and data provided by the installation physical therapy, orthopedic, and podiatry clinics. Fifty-four percent (215) of the women sustained some reportable injury. These injuries resulted in an average training time loss of 13 days. Forty-one percent of these injuries prevented participation in all activity, 31% resulted in only limited participation. The injury data were correlated with prior-fitness measures, documenting that a major cause of injury in women can be attributed to (greater body weight and fat percent), lack of prior conditioning, limited leg strength. These factors, coupled with some inherent physiological characteristics of women, (i.e. wide pelvis, less strength, and greater joint flexibility), contributed to the increased risk of injury in these women. It is concluded that susceptibility to these potential orthopedic and medical conditions can be identified prior to the beginning of training and minimized through proper remedial actions before a strenuous PT program is initiated.

With the rising recruitment of women into the armed forces, data are needed on the their response to physical training and the physical differences that may limit their performance capacity. Although insights have been gained regarding the beneficial effects of physical training on the stamina, muscular strength, and endurance of both men and women in the armed forces [1,2,3] little information is available concerning the risk of injury involved in exposing previously sedentary women to a rigorous physical training program.

In tne past, it has been extremely difficult to study the incidence and distribution of injuries in normal young women because of the relatively small number participating in strenuous physical activity ana the self-selected nature (athletes) of those who do [4,5,6,7]. However, this has changed with the rising interest and participation of women in the whole spectrum of sport activities. The purpose of this prospective research was to 1) determine the incidence and nature of injuries in a female population as a result of a rigorous, supervised endurance training program and 2) identify the predisposing factors that may be related to their occurrence.

METHODS

We followed a group of 400 women recruits, aged 18-29 years old (average age 21) through a complete 8 week basic training cycle (January 15 through March 12, 1978). (See Table 3 for other descriptive characteristics.) A complete medical history was available, and they were given a complete physical examination prior to the beginning of training. Prior to training, an initial assessment of physical work capacity was accomplished. This assessment included the determination of body composition using skinfold estimation and the equation of Durnin Wormseley (8) strength of the major muscle groups, (eg. legs, trunk, arms and upper torso), aerobic capacity (VO_2 max)[9], psychosomatic predisposition using the Health Opinion Survey (HOS),[10] activity history (previous athletic participation), and self-perception of fitness level compared to other women of comparable age prior to beginning training.

These women participated in an integrated (male and female) endurance training and conditioning program 1 hour per day, 5-6 times a week that involved a series of standard warm-up calisthenics including situps, pushups, side straddle hop, leg overs, and modified knee bends. These exercises preceded each training session and progressed from 6 to 12 repetitions per session over the course of the training program. Running began with 3/4 miles a day at a 10 min/mile pace and increased to 2 miles in 18:30 minutes by the end of the 8 week training period. The training program also involved extensive marching and other activities germane to military training.*

For the purposes of this study an injury was defined as any disability which was incurred during or as a result of physical training/conditioning which required attention from the medical facility. Only 327 of the 400 women were available for post training evaluation; 20 additional women were subsequently followed up because they had sustained injuries which required hospitalization. The remaining 53 women were administratively discharged or unavailable for testing, and no information was available on them.

The injury data were gathered through the use of a self report medical disposition questionnaire given to the women following training. These data were supplemented with records from the dispensary, consults and radiographic data provided by the installation physical therapy, orthopedic, and podiatry clinics.

A discriminant function analysis was performed using injury during training

*The training program is outlined in the Drill Sergeant Guide for Pre-Baseline Physical Training. Fort Benning, GA, dtd, 20 Dec 77.

as the criterion variable and using the variables gathered during initial assessment as predictors.

RESULTS

The self report questionnaire data indicated that 54% (215 of 347) of the women had sustained some sort of injury requiring medical attention over the 8 weeks of training. The incidence as tabulated from the questionnaire is presented in Table 1. (It should be noted that the incidence of injury in women compared unfavorably with the incidence reported for men undergoing the same training (26% or 202 of 770). These injuries resulted in an average loss of 13 training days during the basic training cycle, 41% (80) of the injuries prevented participation in all physical activity (major profile) and 31% (61) resulted in limited participation (minor profile). Table 2 presents a summary of the specific diagnoses and structural involvement of these injuries as documented by the hospital consultation sheets and radiographic evidence. As can be seen, the majority were either over use syndrome or stress fx.

Injuries usually resulted from a combination of (1) continued hard training after onset of symptoms, (2) inherent structural weakness, or (3) biomechanical anomaly. It should be noted that tibial and femoral stress fractures accounted for a third of all the injuries identified and represented the most serious sequale of this endurance training program. Figure 1 presents stress fracture data as a function of the onset of symptoms during the 8 week training cycle. The incidence of tibial stress fractures increased during February, and dropped off in March. The hip stress fractures increased throughout training and reached the maximum during the last two weeks of training.

A discriminant function analysis was used to identify the underlying variables that contributed to the prediction of injury. We found that body composition, muscular strength of the legs, previous athletic participation, self perception of fitness and psychosomatic predisposition were correlated with injury (Table 4). The discriminant analysis resulted in a linear combination of the variables that maximally differentiated the two groups. For the present data, (Table 4) the discriminant function is .507 PHY FIT + .683 WT - .662 LEGSTR -.552 PCFAT. However, only 55% of the cases could be correctly classified. Table 5 presents a breakdown of the different degree of injury (major, minor, overuse) within the group of injured women. As can be seen, the variables of PCFAT and LEGSTR are both significantly different across the group and correspond to the previous comparison of injured vs. uninjured women.

DISCUSSION

The data presented document that weight, % body fat, and limited leg strength in women can be attributed to lack of prior fitness or conditioning. These findings also suggest that these factors may be responsible for the increased incidence of injuries occurring in women during training, and any reduction in this incidence is seen as cost effective. Likewise, there are several physiological factors reported in the literature[12] which are generally considered to predispose women to these injuries. The elasticity in the connective tissue, which causes women to be more flexible, may make them more vulnerable to ligament or joint injury. Women's biomechanics and wide pelvis appear to contribute to the increased risk of injury to the hip and the outer aspect of knee, leg and foot because of the varus tilt. This is further aggravated by the apparent lack of heel stability inherent in the Army boot used by the women during basic training. Though the standard Army boot has proved quite satisfactory for men during basic training, women report that the heel width is too great even in the narrow sizes used by them. The resulting heel instability surely aggrevates existant ankle weakness or foot disorder. Likewise, since women are of smaller structure, the 30" step, the very basis of the military drill and ceremony, is often an aggravating factor in the incidence of stress fractures and overuse syndromes. Another factor contributing to injuries in women may be their inability to differentiate between "pushing themselves" beyond the pain threshold and exposing themselves to undue risk of injury. It is evident from the data that many of the symptoms of overuse which occurred early in the training program culminated in injury later on, having been neglected or considered inconsequential initially. Likewise, returning to training before the symptoms had fully disappeared was courting disaster later in training.

A major factor in the development of injuries in this sample is believed to be the rapid onset of training which did not allow for a progressive exposure to stress and the development of tolerance. The more sedentary, unconditioned women were exposed to a greater risk of injury to the lower extremeties when they were put under this physical stress. Initially, the bones attempt to become stronger by remodeling their internal architecture in response to the chronic physical demand. In doing so, they actually become weaker in the area of mechanical stress, and continued training during this period may have lead to injury. The second phase of this remodeling involves actual deposition and hypertrophy of the bone along the lines of stress. However, during the lag time between these two phases, the bone is also more susceptible to fracture[13]. If training had been more progressive some of these injuries might have been avoided.

Stress fractures have been widely studied in men[7] because they can potentially prevent an individual from performing his normal duties for a prolonged period of time. However, with the increasing number of women participating in various physical activities, the occurrence of stress fracture has risen dramatically[6]. These findings were supported by another source of morbidity data (14) which reported that during basic training women had greater than twice the rate of fractures as that reported for men (19.6/1000 compared to 9.4/1000 for women and men respectively). With regards to the causes of stress fractures, the results of Gilbert and Johnson[11] apply equally well to women as to men. The stress fractures are related to body structure and are found to be more common among overweight recruits and those with little exercise experience. However, the majority of the fractures in men reported by Gilbert and Johnson were of the metatarsals and os calcis, or "march fracture", whereas in the present study we found only 8 cases of these types of stress fractures in women. The majority of those found in women recruits were tibial and femoral in nature. A large number of cases of symptomatic chondromalcia of the patella were also reported.

Though the conventional statistical criterion for significance was not achieved, it must be kept in mind that in this case the increased probability of a Type I error ($P \geq .15$) still provides for a substantial improvement (over no information at all) regarding an individual's susceptbility to injury.

CONCLUSION

According to our finding, women tend to have more injuries when they enter an intense training program and these injuries are more severe in nature than those reported for men. The factors that contribute to this increased probability of injury include a lack of physical conditioning prior to entering the program, greater body weight, less leg strength and greater percent body fat than women who did not sustain injuries during the program. However, several other factors also appear to be predisposing factors for injuries in women.

Many women are inexperienced in intense exercise and do not understand "pushing themselves" as a method of developing tolerance (especially earlier in the training cycle); this inexperience may be associated with the risk of subsequent injury. If the training were more progressive, many initial injuries could be avoided. Remedial exercises for the development of lower leg strength and joint stability could also prove to be of value in reducing injury from the clinical standpoint. Likewise, heel orthotics inserted in the boots may reduce injury by stabilizing the foot during running. Probably most important, however, is that personnel must be aware of the signs and symptoms of overuse syndrome and the complication of an untreated stress fracture, especially in the femur or its neck.

Likewise, after monitoring training programs of 3 days a week or less, with only a small number of these injuries (5), we have deduced that training over three days per week results in a significant increase in the injury rate for previously sedentary women. However, the program that maintained a three day/week workout schedule with a day of rest in between, or at least a day of activities that did not involve continual pounding on the legs, had a salutary effect on the incidence of injury in these individuals. It can also be concluded that a soft running surface or shoes designed to absorb the shock of running on hard surfaces would be beneficial in reducing injuries still further.

There can be little doubt that disorders of the lower extremities for the women recruit, like those in men, are costly in terms of medical care and utilization, recruit training time lost, hospitalization and other duty restriction. The solution to the problem is not clear cut because of the multidimen~~s~~tional nature of the problem. Preventative programs such as thorough pre-enlistment screening, that include an assessment of the factors discussed in this paper (eg. prior physical activity, leg strength, body composition and weight), would provide a means of identifying individuals at risk of injury and allow for appropriate action. This could be in the form of remedial physical training and toughening programs, orthotics, and proper breaking in of footwear. Likewise, early identification and treatment of overuse symptoms would be necessary to further reduce the incidence of lower extremity injuries in all recruits entering training, but especially in women because of their increased susceptibility to injury.

REFERENCES

1. Kowal DM, Patton JF, Vogel JA. Psychological states and aerobic fitness of male and female recruits before and after basic training. Aviat Sp Envir Med 49(4):603-606, 1978

2. Kowal DM, Vogel JA, Peterson J. Comparison of strength and endurance training on aerobic power in young women. Med Sci in Sports 9(1):70, 1977

3. Daniels WL, Kowal DM, Vogel JA, et al. Physiological effects of a military training program on males and females. Aviat Sp Envir Med (In press) 1979

4. Pollock ML, Gettman LR, Milesis CA, et al. Effects of frequency and duration of training on attrition and incidence of injury. Med Sci in Sports 9(1):31-36, 1977

5. Tomasi LF, Peterson JA, Pettit G, et al. Women's response to Army training. Phys. Sportsmed 5(6):32-37, 1977

6. Eisenberg I, Allen WC. Injuries in a women's varsity athletic program. Phys Sportsmed 6(3):112-120, 1978

7. Lanham RH. Stress fractures in military personnel. J Amer Podiatry Ass 53:192-195, 1963

8. Durnin JV, Wormesley JW. Body fat assessed from total body density and its estimation from skinfold thickness. Br J Nutr 32:77-92, 1974.

9. Kowal DM, Vogel JA, Patton J., et. al. Evaluation and requirements for fitness upon entry into the Army Proc Sym on Phy Fit NATO Rpt DS/DR (78)98:93-98, 1978

10. McCarroll JA, Kowal DM, Phair PW. The health opinion survey: An exploration of its possible uses in a population of military recruits. J of Health Social Behavior (Submitted) 1979

11. Gilbert RS, Johnson HA. Stress fractures in military recruits - a review of twelve years experience. Mil Med 131:716-721, 1966

12. Goldsmith, NF. Bone mineral in the radius and vertebral asteoporosis in an insured population. J Bone and Joint Surg 55A:1276, 1976

13. Provost, RA, Morris JM. Fatigue fractures of the femoral shaft. J Bone and Joint Surg 57A:487, 1969

14. _________ Health in the Army, 52-53, Jan 1979

ACKNOWLEDGMENT

I would like to express my appreciation for assistance in data collection to Dr. Kish, Podiatry Service, and LTC Ozburn, Physical Therapy Service, Moncrief Army Hospital, Ft. Jackson, South Carolina and to Mrs. L. Kundla and Mrs. D. Caruso for statistical assistance and manuscript preparation.

TABLE 1

SELF REPORT OF THE INCIDENCE OF INJURIES SUSTAINED BY MEN AND WOMEN DURING BASIC TRAINING

	WOMEN (215)		MEN (202)		% DIFF (W-M)
TYPE OF INJURY	N	%	N	%	
FRACTURE (BREAK)	8	3	4	2	1
STRESS FRACTURE	45	21	9	4	17
JOINT PROBLEMS	30	14	47	24	– 10
FOOT PROBLEMS	28	12	54	27	– 15
TENDON INFLAMATIONS	67	31	23	12	19
MUSCLE STRAIN	27	12	26	13	– 1
OTHER	10	4	39	20	– 16

TABLE 2

DIAGNOSED INJURY SUSTAINED BY WOMEN DURING BASIC TRAINING (N = 215)

STRUCTURE	N	%
OVERUSE SYNDROME = (LEG SORENESS, LOWERED GENERAL ENERGY LEVEL, CLUMSINESS AND POOR COORDINATION)	92	42
TIBIAL STRESS FX	45	21
CHONDROMALATIA OF PATELLA	21	9
HIP OR NECK OF FEMUR STRESS FX	20	9
ANKLE SPRAIN	12	6
ACHILLES TENDONITUS	10	4
CALCANEOUS STRESS FX	6	3
ANTERIOR COMPARTMENT & FÁSCIAL STRAIN	6	3
METATARSAL STRESS FX	2	1

TABLE 3

COMPARISON OF SELECTED PARAMETERS FOR INJURED VS UNINJURED WOMEN PRIOR TO PARTICIPATION IN 8 WEEKS OF PHYSICAL TRAINING

VARIABLES	INJURED N = 195	UNINJURED N = 132
BODY WEIGHT (kg) WT	59.2 ± 7.3	59.3 ± 6.8
HEIGHT (cm) HT	162.3 ± 6.8	162.5 ± 6.3
BODY FAT (%) PCFAT	28.4 ± 4.9	27.7 ± 4.4
STATIC STRENGTH OF LEG EXTENSORS LEGSTR (kg OF FORCE)	91.2 ± 32	95.7 ± 28
VO_2 MAX (Ml/kg MIN) VO_2 max	37.9 ± 4.6	36.2 ± 3.3
PREVIOUS ATHLETIC PARTICIPATION (1 VERY INACTIVE - 5 VERY ACTIVE) ATHPAR	3.18 ± .9	3.25 ± .8
PHYSICAL FITNESS COMPARED TO OTHER WOMEN (1 POOR TO 5 SUPERIOR)	2.82 ± .7	2.90 ± .5
HOS SCORE (PSYCHOSOMATIC PREDISPOSITION) HOS	31.3 ± 6.6	30.3 ± 5.5

*$p \leq .05$

TABLE 4

DISCRIMINANT ANALYSIS OF FACTORS CONTRIBUTING TO CLASSIFICATION OF INDIVIDUALS WHO WERE INJURED OR UNINJURED

ACTUAL GROUP	N	PREDICTED GROUP INJURED	PREDICTED GROUP UNINJURED
INJURED	195	52.9	47.1
UNINJURED	132	41.1	58.9

PERCENT OF CASES CORRECTLY CLASSIFIED 55.16%

SUMMARY TABLE OF VARIABLES

STEP NUMBER	VARIABLE ENTERED	APPROX F	RAO'S V	DF	SIGN LEVEL
1	PHY FIT	2.56	2.56	1/295	.10
2	LEGSTR	1.97	3.97	2/294	.13
3	WT	1.55	4.70	3/293	.19
4	PCFAT	1.67	6.74	4/292	.15

TABLE 5

COMPARISON OF PRETRAINING PARAMETERS FOR THE LEVELS OF INJURY SUSTAINED BY WOMEN DURING 8 WEEKS OF PHYSICAL TRAINING (N = 215)

VARIABLES	MAJOR INJURY N = 80	MINOR INJURY N = 66	OVERUSE SYNDROME N = 64
BODY WEIGHT (kg)	59.7 ± 7.4	58.0 ± 6.9	59.8 ± 7.6
HEIGHT (cm)	162.1 ± 6.9	161.3 ± 6.7	163.6 ± 6.9
BODY FAT (%)	29.5 ± 4.4	28.3 ± 5.3	27.4 ± 4.9*
STATIC STRENGTH OF LEG EXTENSORS (kg OF FORCE)	93.5 ± 3.2	92.9 ± 3.2	103.1 ± 3.3*
VO_2 MAX (Ml/kg MIN) VO_2 MAX	38.4 ± 4.2	36.7 ± 4.1	39.2 ± 3.6
PREVIOUS ATHLETIC PARTICIPATION (1 VERY INACTIVE - 5 VERY ACTIVE)	3.18 ± .7	3.19 ± .9	3.20 ± .8
PHYSICAL FITNESS COMPARED TO OTHER WOMEN (1 POOR TO 5 SUPERIOR)	1.85 ± .5	1.91 ± .4	2.10 ± .5
HOS SCORE (PSYCHOSOMATIC PREDISPOSITION)	30.9 ± 6.7	31.6 ± 6.4	31.3 ± 6.7

*$p \leq .05$

CARBOHYDRATE LOADING AS A MEANS OF EXTENDING ENDURANCE PERFORMANCE *

James A. Hodgdon, Harold W. Goforth, Jr.,
and Richard L. Hilderbrand

Naval Health Research Center, P.O. Box 85122, San Diego, CA 92138
Naval Ocean Systems Center, San Diego, CA 92152

INTRODUCTION

There are specialized groups within the military community whose missions require sustained aerobic performance. Within the U.S. Navy such groups include underwater demolition teams (UDT) and sea, air, and land (SEAL) team personnel. UDT and SEAL team missions range from extended beach reconnaissance missions lasting 6-8 hours, to long-range patrols lasting several days. These operations may require walking, swimming, or paddling for long distances and may include climbing and repelling, often while carrying heavy packs. Maximizing the endurance performance capability of these personnel is important in optimizing performance on such missions.

Carbohydrate loading is a popular technique for extending endurance performance, especially among marathon runners. The technique is based on the finding that muscle fatigue appears to be related, at least in part, to exhaustion of glycogen stores in the working muscles, particularly as the work load exceeds 75% of maximum rate of oxygen consumption ($\dot{V}O_2$ max) [2, 5, 9]. Increasing the muscle glycogen stores by manipulation of diet and exercise level (so-called "carbohydrate loading") has been shown to increase the work time to exhaustion on the bicycle ergometer [2] and to decrease the time required to run a 30 km. race [10].

Most commonly a program of carbohydrate loading consists of two phases: a depletion phase wherein the muscles and liver are depleted of their glycogen stores by a combination of strenuous workouts and the intake of a diet low in carbohydrates (CHO); and a loading phase wherein the glycogen stores are increased above their normal level by low intensity workouts and intake of a diet high in CHO [1, 2]. The depletion of glycogen stores increases the activity of glycogen storage mechanisms so that during the loading phase there is a rebound synthesis of muscle and liver glycogen to above the normal level [11, 12]. The highest levels of muscle glycogen storage seem to be achieved when the depletion phase includes both the strenuous exercise and the low-CHO diet, rather than either one alone [13].

As a pilot study to determine whether or not carbohydrate loading can be useful in the military setting, we tested a carbohydrate loading program on UDT and UDT/SEAL personnel from a special warfare group, hypothesizing that

* The results of this work were previously presented at a NATO symposium on Physical Fitness with special reference to Military Forces, D.C.I.E.M., Canada, 3-5 April 1978.

This research was supported in part by the Department of the Navy, Bureau of Medicine and Surgery, under Work Unit No. M0099-PN.004-8050. The views presented in this paper are those of the authors. No endorsement by the Department of the Navy has been given or should be inferred.

such a program would increase their endurance capabilities.

METHODS

Participants

The participants in this study were 9 male UDT and UDT/SEAL personnel, aged 22-36 years. The participants represented a cross-section of the special warfare group in their normal readiness state. Each participant was briefed on the nature of the study and the risks involved in participation in the study. Each participant gave his voluntary consent to participate with the understanding that he could withdraw from the study at any time. Prior to his acceptance into the study, each participant filled out a medical history and was given a physical exam.

Measurements

Baseline measures of age, height, weight, and four skinfolds (biceps, triceps, subscapular and suprailiac) were taken for each participant at the beginning of the study. Skinfolds were measured to the nearest 0.1 mm. with a Harpenden skinfold caliper. Percent body fat was estimated from the total skinfold thickness using the body density equations of Durnin and Womersley [7] and the density to percent body fat conversion of Siri [14]. In addition, the $\dot{V}O_2$ max was determined for each participant using a discontinuous treadmill-running protocol [15].

$\dot{V}O_2$ was determined by open circuit spirometry using a system composed of an O_2 analyzer, a CO_2 analyzer, a spirometer, and an air temperature thermometer interfaced with a programmable desk-top calculator. The participant breathed through a Daniel's valve and a sample of the expired gas was drawn off and analyzed on-line. At 15-second intervals $\dot{V}O_2$ values were calculated and printed out. The $\dot{V}O_2$ value is taken to be the average of the last 1 minute of gas collection.

The participants were rank-ordered by their $\dot{V}O_2$ max values and assigned to one of two experimental groups in alternate order by rank. The number of participants ranked was reduced by attrition from 24 volunteers to 9 who completed the study successfully. The relevant physical and physiological characteristics of the participants completing the study are given in Table I. The two group means do not differ significantly for any characteristic ($p > 0.05$, t-test comparison of independent means).

The hypothesis that carbohydrate loading increases endurance performance was tested by comparing endurance performance following a carbohydrate-loading program with performance following a non-loading program. Each participant followed each program one time and thus served as his own control. A counter-balanced experimental design was used, the order of treatment (loading vs. non-loading programs) was reversed for the two experimental groups. The hypothesis was tested using a t-test for planned comparisons among sample means [8]. The null hypothesis was to be rejected at $p \leq 0.05$.

Table I. CHARACTERISTICS OF THE PARTICIPANTS

	Participant No.	Age (yrs)	Ht. (cm)	Wt. (kg.)	% body fat	$\dot{V}O_2$ max (l/min)	$\dot{V}O_2$ max (ml/kg.min)
Group I							
	06	32	175.3	76.5	14.9	5.43	71.11
	22	36	182.9	83.5	18.2	4.47	53.48
	28	30	170.2	60.8	12.6	3.19	52.43
	42	25	172.7	79.4	16.0	4.93	62.03
	86	30	182.9	71.7	12.1	4.73	66.01
$\bar{X}$		30.6	176.8	74.4	14.8	4.55	61.01
SD		(4.0)	(5.9)	(8.7)	(2.5)	(0.84)	(8.04)
Group II							
	11	25	182.9	74.9	10.8	5.30	70.73
	51	22	188.0	79.9	6.7	4.19	52.46
	69	30	162.6	67.6	16.7	4.31	63.66
	80	26	170.2	64.0	10.7	4.19	65.45
$\bar{X}$		25.8	175.9	71.6	11.2	4.50	63.08
SD		(3.3)	(11.6)	(7.2)	(4.1)	(0.54)	(7.69)
Sample $\bar{X}$		28	176.4	73.1	13.2	4.53	61.93
SD		(4.3)	(8.2)	(7.7)	(3.6)	(0.68)	(7.46)

Endurance performance was measured as the length of time a participant could run on a motor-driven treadmill at 0% grade at a speed requiring him to work at approximately 80% of his $\dot{V}O_2$ max. The endurance run was conducted in an interrupted fashion. The participant would run for 18 minutes and would then be allowed a 2 min. rest. Water was provided ad libitum during the run. Electrocardiogram, heart rate, and rectal temperature were monitored throughout the run. Near the midpoint of each 18-min. running period, $\dot{V}O_2$ was determined by open circuit spirometry. The treadmill speed was adjusted to maintain a work load of approximately 80% $\dot{V}O_2$ max. The test was terminated when the participant indicated he could not continue running at this work load. For the second endurance test, the speed/time profile of the first test was repeated.

Diet/Exercise Programs

Following the suggestions of Åstrand [1] and Karlsson and Saltin [10], a 6-day loading program was used. The program is described in Table II. During the depletion phase, the participant ran 14 miles on Day 1, 6 miles on Day 2, and 4 miles on Day 3. The first day's run was more than 3 times his normal workout (∿ 4 miles), and was performed to deplete the running muscles of glycogen. On the average, the depletion run required 111.8 minutes to complete. The runs on Days 2 and 3 were intended to keep the muscles glycogen-depleted. The diet for the depletion phase was one which was low in carbohydrate. During the loading phase, the participants ran only short distances with no running exercise on the day prior to the endurance test. The diet for this phase was high in CHO to promote glycogen synthesis and storage. On Day 6 they were given a normal composition diet.

As a control, non-loading program, each participant was asked to run 4 miles each day (about equal to their normal workout), except for the day before the endurance test when they did not run. During these 6 days, they ate a diet of approximately normal composition in terms of calorie percentages of fat, protein, and carbohydrate [4]. This program is described in Table III.

Food for the participants was provided by the experimenters. The diet consisted of a combination of solid food and a liquid formula. The amount and kind of solid food was the same for the three diets (Low-CHO, High-CHO, and normal). The proportions of the constituents of the liquid formula varied with each particular diet. The composition of the diet and the proportion of the calories from CHO, fat, and protein are given in Table IV. The Low-CHO diet contains 32g CHO; the High-CHO diet 533g and the "normal" diet 353g. The daily caloric intake provided was 3500 kilocalories.

RESULTS

The results are presented in Table V. The running times presented do not include the rest periods. On the average, the participants ran 10.8 minutes longer following the loading program. This difference represents a 9.0% increase over the nonloaded running time and was significant ($t = 2.43$, $p < 0.05$). There was one participant who had a shorter running time

Table II. CARBOHYDRATE LOADING DIET/EXERCISE PROGRAM

	Day	Diet	Exercise
Depletion Phase	1	Low CHO	14-mile run
	2	Low CHO	6-mile run
	3	Low CHO	4-mile run
Loading Phase	4	Hi CHO	1-mile run
	5	Hi CHO	1-mile run
	6	Normal	No run
Test	7	Normal	Endurance run

Table III. NON-LOADING DIET/EXERCISE Program

	Day	Diet	Exercise
	1	Normal	4-mile run
	2	Normal	4-mile run
	3	Normal	4-mile run
	4	Normal	4-mile run
	5	Normal	4-mile run
	6	Normal	No run
Test	7	Normal	Endurance run

Table IV. DIET COMPOSITION

Solid Food:*

6.5 oz. water-packed tuna
5.0 oz. canned chicken
2 hard boiled eggs
50 g mayonnaise
lettuce (ad libitum)

*constant for all diets

Liquid Formula:**

Calcium caseinate
Corn oil
Fructose
Polycose
Minerals, flavorings and saccharin
Approximately 1900 ml water

**proportions vary with diet

Fluids:

Water, diet soda, and coffee provided ad libitum

% Calories From:	Low-CHO Diet	High-CHO Diet	Normal Diet
CHO	3	64	46
Fat	50	24	42
Protein	47	12	12

Total Calories = 3500/day

Table V. RESULTS

	Participant No.	1st Endurance Run Time, t_1 (min)	2nd Endurance Run Time, t_2 (min)	Load Time - Nonload Time (min)
Group I				
	06	82.75 *	73.50	9.25
	22	132.00	129.75	2.25
	28	156.00	163.00	-7.00
	42	138.25	124.00	14.25
	86	149.75	112.50	37.50
$\bar{X}$		131.75	120.55	11.25
SD		(28.96)	(32.32)	(16.70)
Group II				
	11	119.00	127.50 *	8.50
	51	139.25	149.50	10.25
	69	106.50	124.50	18.00
	80	113.50	117.50	4.00
$\bar{X}$		119.56	129.75	10.19
SD		(14.09)	(13.82)	(5.84)

~~~~~~~~~~~~~~~~~~~~~~~~~~

Mean Running Time (min)

Loaded	130.86	(22.18)
Not Loaded:	120.11	(24.43)

Mean Time Difference (min)

Due to Loading:	10.78
Due to Experience (t_1-t_2):	0.50
Due to Groups (G_1-G_2):	1.50

* CHO-loading program
~~~~~~~~~~~~~~~~~~~~~~~~~~

following the loading program than he did following the nonloading program.

There was essentially no difference in mean running times due to repeated testing. The mean running time was only 0.5 min. longer for the first run than for thc second. The mean difference in running time between groups was only 1.5 min. (group I running longer). These were not significant effects.

Spearman rank correlation coefficients were computed for the difference in running times between the two programs and each of the physical characteristics given in Table I. This was done to determine whether or not physical makeup or aerobic fitness influenced the effectiveness of the loading program. In no case was a significant correlation found.

DISCUSSION

Our results demonstrate that a carbohydrate loading program can significantly improve endurance performance. Our findings are in agreement with those reported by Karlsson and Saltin [10] who found a 5.4% (7.7 min.) mean decrease in the time required to run a 30 km race when the runners followed a CHO-loading program similar to the one used in this study. The results also are in agreement with Bergström _et al_ [2]. They found the work time to exhaustion on a bicycle ergometer at 75% of VO_2 max increased by 46.5% (52.9 min.) on the average following a carbohydrate diet when compared to the work time following a normal mixed diet. In their study, the exercise and diet schedules were somewhat different from ours, but did involve a glycogen depletion followed by a CHO load.

A difference between our CHO-loading program and others reported is we supplied a normal-composition diet on the last day of the loading phase of the CHO-loading program, rather than a high-CHO diet. This was done to eliminate the intestinal disturbance which may result from an intake of high-CHO meals (unpublished survey of marathon runners, Goforth 1976). The data of Saltin and Hermansen [13], and Piehl [15], indicate the rise in muscle glycogen is essentially complete within 2 days and that the maintenance of a high-CHO diet for the third day does not appreciably change the muscle glycogen content. Additionally, Bergström and Hultman [14] report that, in non-exercising persons, feeding a high-CHO diet, a low-CHO diet, and even fasting had only minor effects on muscle glycogen levels. The 353g of CHO provided in the normal diet is more than is required to meet normal metabolic needs [4]. We do not feel, therefore, feeding a normal diet on day 6 changed the effectiveness of the load.

A common finding in studies of CHO-loading is the large individual variation in response to the CHO-loading program. In the present study we find differences in endurance time (i.e., loaded-nonloaded) ranging from -7 minutes (a 6% decrement) to 37.5 minutes (a 33% increase). Similar variation in performance change with loading is seen in the studies of Karlsson and Saltin [10] and Bergström _et al_ [2]. The lack of significant correlation between our measured physical characteristics and change in performance with loading suggests that physical makeup and aerobic fitness as we measured them are not a meaningful source of variation. Similar lack of correlation is found when we analyze the data presented by Karlsson and Saltin [10] and Bergström et al [2]. However, factors such as proportion of the various muscle fiber types were not measured. It is primarily the high oxidative fibers which are recruited in a run of this intensity [5]. Thus individual differences in the proportion of high oxidative fibers implies differences in the number of fibers which, when loaded, can contribute to augmentation of performance.

Also in our study the intake of coffee and other caffeinated beverages was not controlled. The work of Costill et al [6], published after completion of our study, shows augmentation of endurance performance with caffeine ingestion. This finding suggests that variation in loading effectiveness could result from variation in pre-run coffee drinking patterns between each of the runs.

Other factors which need to be considered in explaining the variation in effectiveness are differences in motivational levels both between persons and temporally within person, and individual variation in metabolic response to the loading program. The sources and degree of individual variation must be identified and understood if a CHO-loading program for military application is to be developed.

Additional research is needed before the efficacy of a field-applied program can be determined. It is difficult to make an extrapolation from performance gain on a laboratory task at a fixed work rate to performance enhancement on a military mission. Therefore one requirement for further work is the development of a more militarily-relevant criterion task, perhaps an obstacle course or a practice mission scenario.

Also, it seems likely that troops will not always have six days to prepare for combat missions. Therefore every effort must be made to shorten the loading time. As mentioned above, it appears the loading phase can be shortened to two days [13, 15]. The depletion might be similarly shortened.

Troops will not always have the space available to do depletion runs. Therefore the effectiveness of alternative exercises such as running in place must be explored. Finally, the relative contributions of the diet and exercise portions of the loading program to its effectiveness must be assessed, so that the program can be optimized for a variety or possible time and space constraints.

SUMMARY

This study was designed as a pilot to test the effectiveness of a CHO-loading program on a cross-section of a military population which might benefit operationally from such a program. We have found a significant increase in the length of time our participants could run at 80% of $\dot{V}O_2$ max before reaching exhaustion. Further work is needed before the effectiveness of such a program in the field setting can be determined.

REFERENCES

[1] Åstrand, P-O. 1967. Diet and athletic performance. Fed. Proc. 26(6): 1772-1777.

[2] Bergström, J., L. Hermansen, E. Hultman and B. Saltin. 1967. Diet, muscle glycogen, and physical performance. Acta physiol. scand. 71:140-150.

[3] Bergström, J. and E. Hultman. 1972. Nutrition for maximal sports performance. J. Amer. Med. Assoc. 221:999-1006.

[4] Calloway, D. H. 1971. Dietary components that yield energy. Environ. Biol. Med. 1:175-186.

[5] Costill, D. 1974. Muscular exhaustion during distance running. Phys. and Sports Med. 2:36-41.

[6] Costill, D. L., G. P. Dalsky and W. J. Fink. 1978. Effects of caffeine ingestion on metabolism and exercise performance. Med. Sci. Sports. 10(3):155-158.

[7] Durnin, J. V. G. A. and J. Womersley. 1974. Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. Br. J. Nutr. 32:77-97.

[8] Hays, W. L. 1963. Statistics for Psychologists. New York, Holt, Rinehart, and Winston, pp. 462-466.

[9] Hermansen, L., E. Hultman and B. Saltin. 1967. Muscle glycogen during prolonged severe exercise. Acta physiol. scand. 71:129-139.

[10] Karlsson, J. and B. Saltin. 1971. Diet, muscle glycogen, and endurance performance. J. Appl. Physiol. 31:203-206.

[11] Lamb, D. R., J. B. Peter, R. N. Jeffress and H. A. Wallace. 1969. Glycogen hexokinase and glycogen synthetase adaptations to exercise. Amer. J. Physiol. 217:1628-1632.

[12] Piehl, K., S. Adolfsson, and K. Nazar. 1974. Glycogen storage and glycogen synthetase activity in trained and untrained muscle of man. Acta physiol. scand. 90:779-788.

[13] Saltin, B. and L. Hermansen. 1967. Glycogen stores and prolonged severe exercise. In Blix, 6. (Ed.) Nutrition and Physical Activity. Uppsala, Sweden, Almqvist and Wiksells.

[14] Siri, W. G. 1961. Body composition from fluid spaces and density: Analysis of methods. pp. 223-244. In Brozek, J. and A. Henschel (Ed.) Techniques for Measuring Body Composition, Washington, D. C., National Academy of Sciences, National Research Council.

[15] Taylor, H. L., E. Buskirk and A. Henschel. 1955. Maximum oxygen intake as an objective measure of cardio-respiratory performance. J. Appl. Physiol. 8:73-80.

EXECUTIVE ASSESSMENT CENTERS

IN THE PUBLIC SECTOR

Prepared by

Burton F. Krain, Ph.D.
U.S. Office of Personnel Management
Great Lakes Region

Presented to

Military Testing Association

Annual Conference
San Diego, California
October 17, 1979

Executive Assessment Centers in the Public Sector

Two recent developments have strongly influenced the growth of Assessment Centers in the public sector. The issuance of the Uniform Selection Guidelines in August, 1978 has placed greater emphasis on agencies in the public sector to document their selection procedures and merit promotion practices where adverse impact is found to exist. The second development is the passage of the Civil Service Reform Act in October, 1978. This Act has placed greater responsibility on the newly formed U.S. Office of Personnel Management to provide consulting services to federal, State and local sector agencies.

The Great Lakes Region of the U.S. Office of Personnel Management is unique insofar as it has a cadre of highly experienced personnel psycnologists with experience in public sector assessment center technology. With over one million dollars of resources flowing to State and local governments in assessment centers over the past few years as a result of Intergovernmental Personnel Act (IPA) grant funds, the Great Lakes Region has experienced every variation on assessment centers that the state-of-the-art allows. These experiences include assessment centers for upward mobility, selection, placement, merit promotion and executive development. Our experiences with the Air Force, Department of Agriculture, U.S. Forest Service and the Department of the Army has led us to some practical conclusions about the future of assessment centers in the public sector.

First, there most certainly is a future for assessment centers in the public sector. Documentation requirements under FPM Supplements' 271-1, 271-2 and 335-1 mandate a systematic approach to selection and merit promotion. The assessment center process is such an approach. The established validity and reliability of the process has been demonstrated repeatedly in both public and private sector studies. As a recent article in Business Week reports (October 8, 1979): "today there are more than 2,000 corporate-run assessment centers, up from no more than 100 a decade ago." The same trend is true in the public sector as well.

The second conclusion about assessment centers is that for them to continue to expand in the future they will have to improve on their cost-effectiveness. Douglas Bray from AT&T was mentioned in the same Business Week article as "experimenting with assessment techniques that can be administered more cheaply to individual candidates." Our practical experience in the administration of assessment centers is that their can be efficiencies introduced into the process without sacrificing the quality of the overall process.

These efficiencies include more directive training to the exercises used and how they work in the assessment process. The use of core assessor training (training of a general nature) and situation-specific assessor training (exercise specific) can reduce overall training time, especially with assessors previously trained from other centers. Besides more directive training, the actual assessment process can be shortened from several days to one day with no apparent loss of quality in the final decisions. This is done through tailoring the exercises via a thorough job analysis rather than through the use of off-the-shelf exercises.

Our practical experience also dictates that the assessment center process receives greater credibility and acceptance by candidates when exercises are perceived as job-related and situation specific. It is for this reason as well as for meeting the requirements of the Uniform Selection Guidelines that job analysis receives so much emphasis in our assessment center projects. Dr. Marilyn Hafer, formerly a personnel psychologist with the U.S. Office of Personnel Management will lead a discussion on streamlined techniques in job analysis for assessment center development. These techniques, developed in cooperation with the Department of the Army will result in considerable cost reductions for future assessment centers job analysis studies. So, with greater emphasis on assessor training that is situation specific and job-related and with greater emphasis on job analysis and supportive documentation that is collected in a structured manner, we view the outlook for assessment centers in the public sector to be optimistic.

In an effort to meet the responsibilities of consultative assistance the U.S. Office of Personnel Management initiated a model consultative services project with the U.S. Army Armament Materiels Readiness Command (ARRCOM) at Rock Island Arsenal, Rock Island, Illinois in the Fall of 1978.

The Assessment Center for ARRCOM at Rock Island was initially developed for executive development purposes. It has since been modified for use in their merit promotion program. The project is the first in a series of consultative services efforts undertaken by the U.S. Office of Personnel Management, Great Lakes Region to ensure the soundness of selection procedures and executive development practices in the agencies it services. The project was successful because it incorporated the aspects I have just mentioned in its initial design. The results of the center will have profound impact on the career paths of the top executives in ARRCOM because they became integrated into the project through feedback on their performance and additional training so that they may serve as an assessor themselves in future assessment center initiatives. Dan Naert, the key resource person from Rock Island Arsenal on Assessment Centers will now explain the unique adaptation of the ARRCOM Assessment Center for Executive Development. Marilyn Hafer will then explain how job analysis techniques were adapted for ARRCOM assessment center developmental purposes. When we conclude these presentations we will open up the discussion, to you, the audiencefor questions or comments.

Rock Island Arsenal Training and Development Division's Role in the Assessment Center Process

by

Dan Naert

In February 1977, Rock Island Arsenal began to accumulate information on the Assessment Center process.

At the time, the Army's Civilian Executive Development Program existed but assumed a passive position. Guidance was provided on how to identify a manager but little effort was made to provide an effective procedure to identify high potential employees who would later serve in managerial positions. Development of these employees is a Department of Army primary goal for its Executive Development Program and a prime source for filling future managerial positions.

The Assessment Center Process was observed and looked upon as a vehicle to identify high potential employees who would be trained to be competitive for managerial positions.

After the Arsenal secured a background of Assessment Center knowledges, a contract to conduct the first Assessment Center was awarded to the Office of Personnel Management, Great Lakes Region. This first Assessment Center was completed in January 1978. This Center involved thirteen Incumbent Managers and nine Civilian Personnel office supervisors. All assessed individuals were employees of Rock Island Arsenal.

This Assessment Center provided information to the Commander and executive management personnel which was relatively known but never before was an outsider or process able to provide job-related data so accurately, realistic and in such a short period of time.

During this time frame, two executive positions were being filled. At the Commander's request, the Assessment Center process was used as a selection assist tool in filling the positions. These positions were the HQ ARRCOM Comptroller, GS-16 and the Rock Island Arsenal Quality Assurance Director, GS-14.

In order to provide the Rock Island Arsenal Assessment Center Administrator "hands on" experience, a developmental Assessment Center was conducted. This center involved six key participants. Later, the process was used as a selection assist in filling the position, HQ ARRCOM Materiel Management Director, GS-15.

Based on the successes of the Rock Island Arsenal project, the Commanding General requested that an Assessment Center model be created for HQ, ARRCOM. Based on the professional services provided by the Office of Personnel Management, Great Lakes Region, this Office was requested to provide contractural services. The project was started in September of 1978 and completed according to the planned schedule in August of 1979.

The HQ ARRCOM Assessment Center model involved twenty-seven incumbent managers who participated as assessees and who later were trained as assessors.

At the present time, Rock Island Arsenal/HQ ARRCOM have an Assessment Center Model for each activity. Managerial personnel and/or Civilian Personnel Office Supervisors have been professionally trained to serve as assessors for future Assessment Center programs.

Since the conception of the Assessment Center process at this installation, ten Assessment Center programs have been conducted. Five of the Centers have been for selection assist and five were conducted for developmental purposes. At the present time, an Assessment Center is pending for a selection assist purpose in filling the HQ, ARRCOM Equal Employment Opportunity Officer position.

Following are benefits derived from the Rock Island Arsenal/HQ ARRCOM Assessment Center Projects:

DEVELOPMENTAL ASSESSMENT CENTERS

1. Profiled individual manager's position through Job Analysis.

2. Ascertained relative importance of managerial dimensions through Job Analysis.

3. Provided managers with additional observation skills in terms of observing human behavior.

4. Provided managers insights and reinforcement of their own strengths and weaknesses and those of their peers.

5. Provided managers knowledge and information on how to improve their Individual Development Plans.

6. Provided managers a tool to accomplish a more objective approach relative to conducting performance appraisals for their subordinates.

SELECTION ASSIST ASSESSMENT CENTERS

1. Conducted a thorough job analysis for Key Managerial positions.

2. Profiled strengths and weaknesses relative to the current incumbent of managerial positions.

3. Tailored exercises around job analysis information.

4. Assisted in the selection process of five key managerial positions. Individuals who were highly rated by the Assessment Center were selected.

ASSESSMENT CENTER DIMENSION DEFINITIONS

Leadership and Supervision

Ability to guide and direct others and initiate action to influence events. Constructively uses influence and persuasion to gain conformance with or without authority. Ability to motivate, appraise and assist others in job accomplishment and career development.

Communication Skills

Ability to clearly and concisely express ideas in written and oral form in group and personal presentations.

Planning and Organization

Ability to independently establish clear cut attainable objectives and forsee resources needed for task accomplishment (men, money and materials). Ability to establish and/or quickly adjust priorities to attain objectives and meet unique problems, changing situations and short deadlines. Ability to effectively coordinate simultaneous assignments for self and others.

Reasoning and Analytical Ability

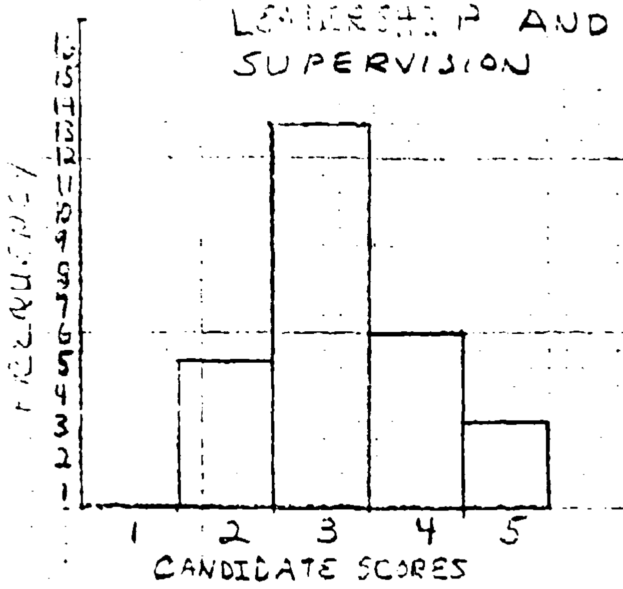
Ability to draw inferences or conclusions from known or assumed facts to identify existing and potential problems and possible solutions. Ability to accurately interpret and apply written and oral instructions and regulations. Skill in analyzing problems carefully and logically by taking all relevant information into account.

Interpersonal Skill

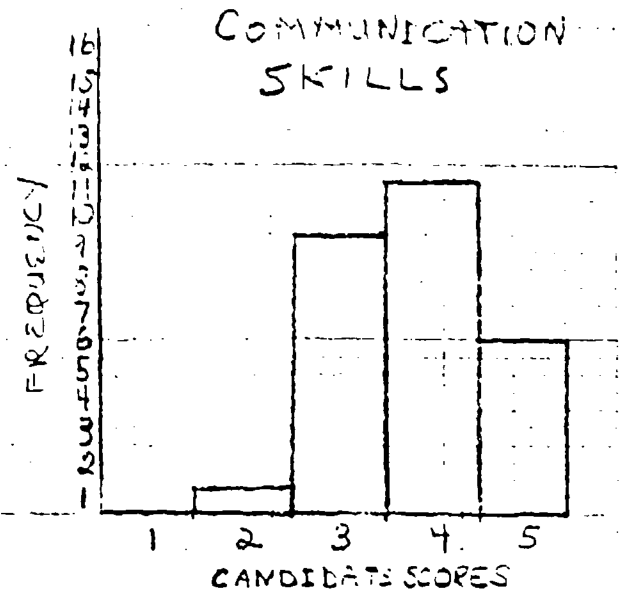
Ability to perceive and react to the needs of others and to accurately perceive one's impact on them. Skill in both offering and gaining cooperation from those inside and outside of the Directorate. Displays tact and diplomacy in all interactions.

Judgment and Decision Making

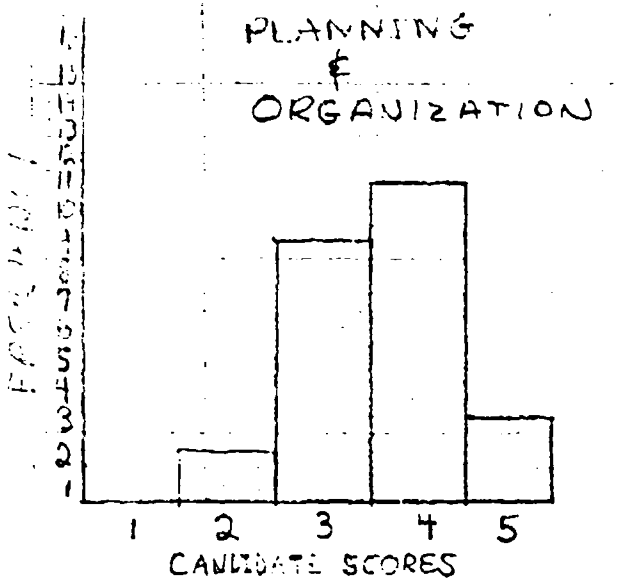
Appropriately decides when to act independently and when to inform or refer to higher authority. Ability to evaluate alternatives and choose the most appropriate course of action. Ability to deal with ambiguous problems, make immediate decisions and establish specific courses of action.



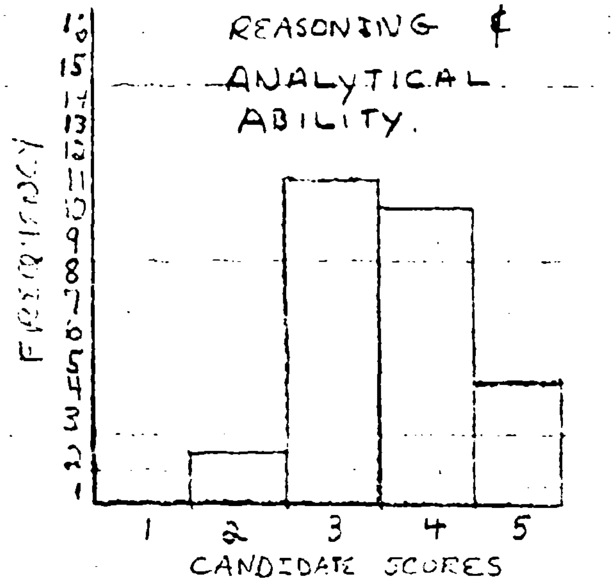
LEADERSHIP AND SUPERVISION



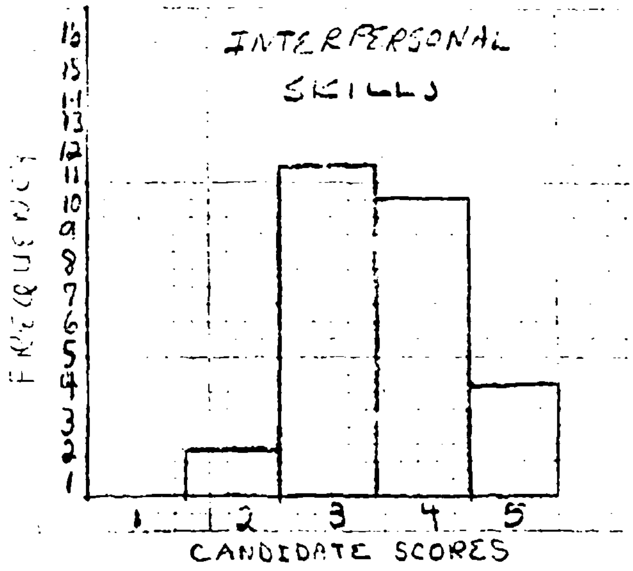
COMMUNICATION SKILLS



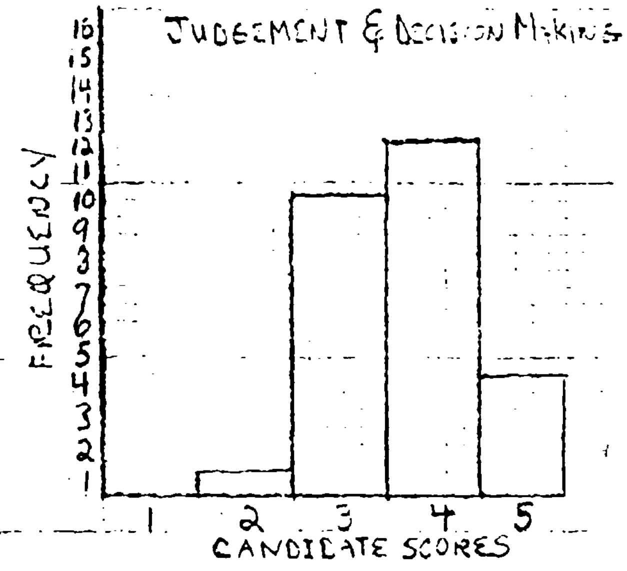
PLANNING & ORGANIZATION



REASONING & ANALYTICAL ABILITY.

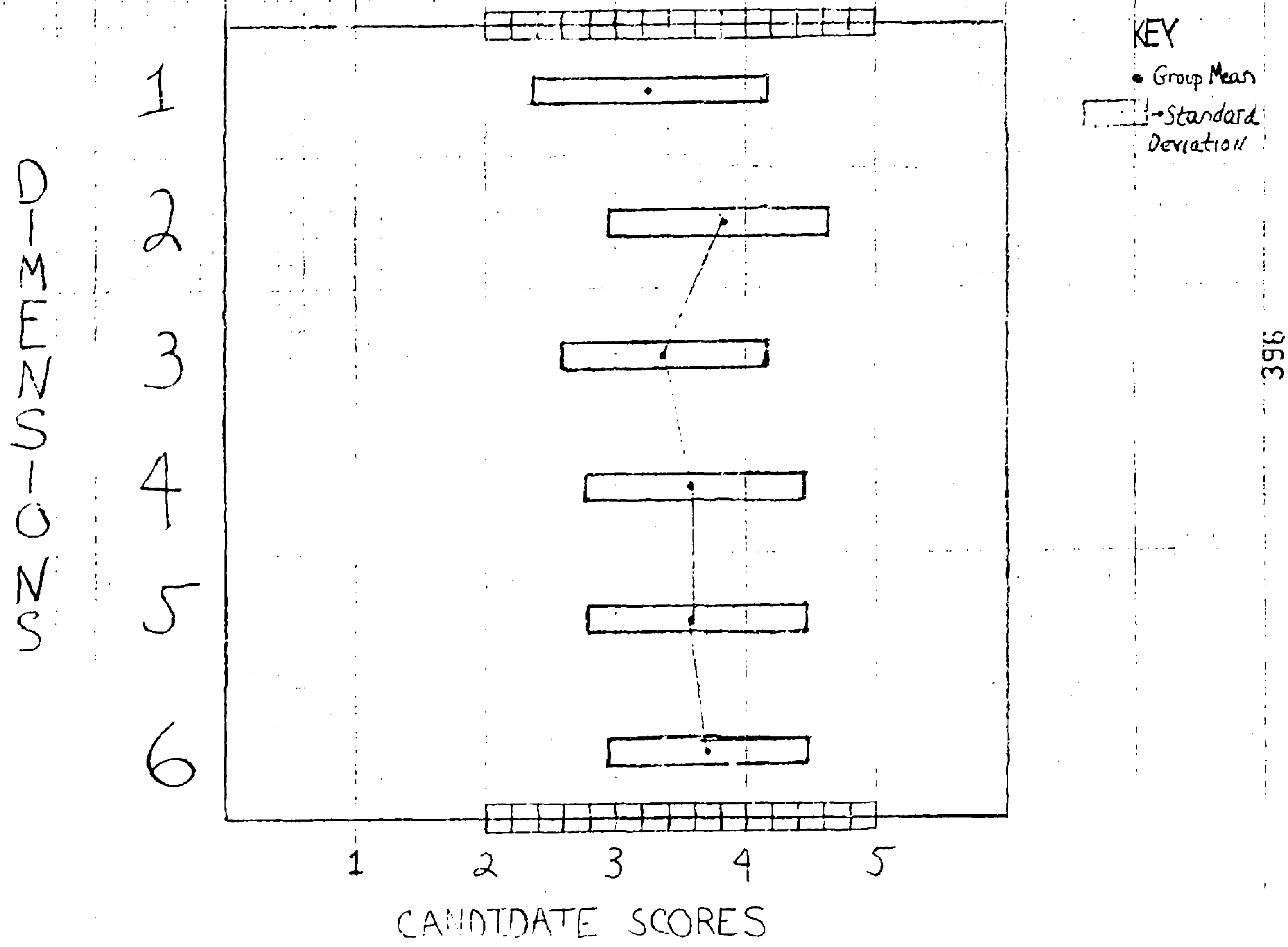


INTERPERSONAL SKILLS



JUDGEMENT & DECISION MAKING

HQ ARRCOM ASSESSMENT CENTER
SUMMARY PROFILE



In addition to conducting the Assessment Center Program, the Office of Personne[illegible]gement administered a Managerial Training Needs Profile. The information for this program was provide[illegible]o each manager prior to the day of assessment and was completed on the assessee's own time. The data for the profile was computer scored and the group itself was used as the norming sample. All outcome measures are relative to this specific group. Following are the results of this profile:

ITEM	ACTIVITY	NO. MGRS WITH ABILITY BELOW REQUIRED	NO. MGRS WITH ABILITY EQUAL TO REQUIRED	NO. MGRS WITH ABILITY ABOVE REQUIRED	TRAINING PRIORITY
14	Public Relations. Acts involving (a) the dissemination of information to persons outside the immediate unit (e.g., supervisor, persons in other units, the general public) concerning the functions and activities of the unit, and (b) the creation of favorable attitudes among these constituencies toward the unit or the organization in general.	12 (48%)	7 (28%)	5 (20%)	1.041
1	Planning. Acts involving (a) the establishment of short- and long-range goals, objectives, and priorities for the work unit with respect to the type quality, or volume of product or service to be provided, and (b) the identification of resources needed to carry out those programs and objectives.	8 (32%)	12 (48%)	4 (16%)	0.979
6	Interpretation. Acts involving the explanation and clarification of directives, regulations, plans, policies, practices, and procedures to give them meaning to superiors, colleagues, and subordinate personnel and to establish their relevance to the tasks these individuals face.	12 (48%)	6 (24%)	6 (24%)	0.970
7	Technical Supervision. Acts involving the direct guidance, motivation, and control of subordinate personnel in the performance of their daily activities.	10 (40%)	8 (32%)	6 (24%)	0.752

ITEM	ACTIVITY	NO. MGRS WITH ABILITY BELOW REQUIRED	NO. MGRS WITH ABILITY EQUAL TO REQUIRED	NO. MGRS WITH ABILITY ABOVE REQUIRED	TRAINING PRIORITY
2	Coordination. Acts involving the integration of activities of individuals within the work unit (or the integration of the activities of members of the unit with those of other organizational units) to reduce duplication of effort, eliminate overlaps of responsibility, and improve the economy and control of operations.	9 (36%)	8 (32%)	6 (24%)	0.752

Knowledges in which training is required by 50% or more of the managers:

EQUAL EMPLOYMENT OPPORTUNITY/AFFIRMATIVE ACTION PROGRAMS--legal constraints and requirements affecting the selection, promotion, transfer, and the dismissal of employees.

CONFLICT MANAGEMENT TECHNIQUES--the cause of conflict and techniques for conflict resolution.

LONG-RANGE PLANNING TECHNIQUES--techniques for determining the long-range needs of an organization and the resources required to meet those needs.

EFFECTIVE READING PRACTICES--how to read fast, with high levels of comprehension.

ORGANIZATION DEVELOPMENT STRATEGIES AND TECHNIQUES--long-range strategies and techniques for improving organizational performance and health.

MANAGERIAL PSYCHOLOGY--applications of psychological concepts, techniques, approaches, and methods towards increasing the effectiveness of managers and organizations.

Abilities in which training is required by 50% or more of the managers:

Be able to make effective use of time.

Be able to analyze and determine probable causes of complex problems.

Be able to ask clear questions and elicit desired information.

Be able to interpret proposals or actions in the context of the goals and values of others.

Be able to provide negative feedback to others in a manner which does not arouse defensiveness or animosity.

Be able to accurately assess how others see you.

JOB ELEMENT APPROACH TO DEFINITION OF ASSESSMENT CENTER DIMENSIONS

Marilyn Hafer Ph.d.

Southern Illinois University
Carbondale. Illinois 62901

INTRODUCTION

One of the issues to be settled in establishing an assessment center is the nature of the dimensions of performance to be assessed. While there are similarities in the lists of assessment dimensions used by various organizations, the process of identifying these dimensions is important. Analysis of the behaviors of managers to establish the job relatedness of specific dimensions is required (Jeswald, 1977). This process provides a basis for validity of the measures obtained from an assessment center, and meets the standards set by the uniform guidelines on employee selection procedures (U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978).

Various techniques have been used in the development of assessment dimensions. Byham (1970) suggests that discussions among managers familiar with target jobs be used to derive dimensions; others (e.g., Slivinski & Desbians, 1970) have used a mail survey consisting of a questionnaire in standardized form, or the critical incident technique (Flanagan, 1954).

In developing an assessment center (designed primarily for developmental purposes) for high-level civilian managers of the U.S. Army Armament Readiness Command at Rock Island, Illinois, a combination of methods was used to define dimensions. Interviews were conducted with incumbent managers following a standardized questionnaire format which included reports of critical incidents, and various paper and pencil measures were obtained at the conclusion of each interview. The primary method of identifying dimensions to be assessed was based upon Primoff's (1975) job element approach.

METHOD

Subjects

Twenty-six male and one female ARRCOM managers were interviewed regarding job requirements. The average age of these managers was 52, and they reported a mean of 20 years managerial experience. Seventeen managers reported attainment of Bachelor's or advanced degrees, primarily in the fields of engineering or business administration.

Instrument

The lists of 135 elements identified as applicable to many jobs (Primoff, 1975), plus one additional item regarding knowledge of equal employment opportunity were given to each manager interviewed. S/he was instructed to pick out 20 items using a job element blank form (Figure 1). The first column on the job element form allows respondents to record the item numbers selected for rating. In columns 2-5, ratings for each item are recorded as plus, check or zero (+ ✓ 0), successive lines allow recording of responses for each item. Managers were instructed to consider a barely acceptable worker in his/her position, and record a plus in column 2 (B) if all would have the knowledge, skill, ability, trait or characteristic indicated for that element, to check if some have it, or record zero if almost none would have it. In column 3 (S), plus was assigned to an element which would be very important "to pick out superior workers," check indicated valuable, and zero indicated the element does not differentiate. The third column (T) was marked to indicate "trouble likely if (the element is) not considered," with plus indicating much trouble; check some trouble, and zero safe to ignore. The fifth column relates to practicality (P). Managers were instructed to consider whether all job openings, some openings, or almost no openings for positions similar to their own could be filled if this element were demanded of candidates, again using the 3-point rating system to indicate responses. The remaining columns of the job element blank are used in hand calculation of values assigned by respondents.

Procedure

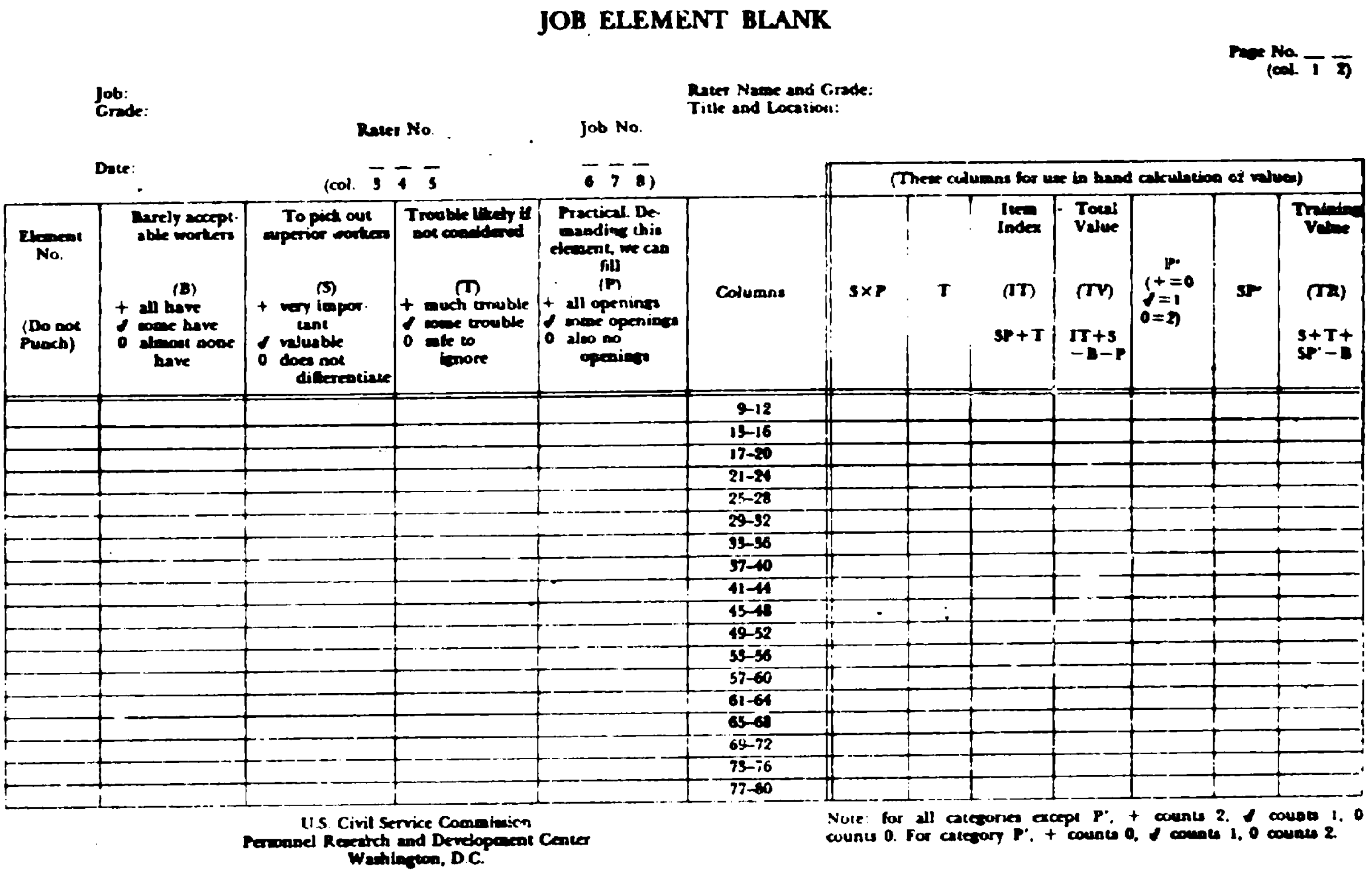
Interviews were conducted by U.S. Office of Personnel Management staff (three psychologists and two personnel specialists). Two interviewers were typically assigned to each manager, although five interviews were conducted by either one or three persons. Time required for interviews ranged from 2 1/2 to 4 1/2 hours. Each manager completed the job element form following the interview.

Analysis

Each manager's choices of items from the job element blank and ratings of those items were transcribed for analysis. All managers did not rate every item chosen, so only scorable responses were used in calculating summary scores. Values of 0, 1 or 2 were assigned to the zero, check or plus ratings, respectively. For each item rated by three or more managers, a summary score was obtained for that element for each response category (B, S, T, or P). Thus for any given element, the managers rating that element were considered to comprise a panel of raters. Summary scores were used in calculating transmuted values, following Primoff's (1975) method (See Appendix A for calculation of transmuted values).

RESULTS

Table 1 shows the results of the transmuted group sums calculated for the four response categories, and three additional values which represent an item's total value as an element, its item index, and the training value of an item. Items chosen from the job element list are shown (Table 1) in



JOB ELEMENT BLANK

Page No. __ __
(col. 1 2)

Job:
Grade:

Rater Name and Grade:
Title and Location:

Rater No. __ __ __ (col. 3 4 5)

Job No. __ __ __ (6 7 8)

Date:

						(These columns for use in hand calculation of values)						
Element No. (Do not Punch)	Barely acceptable workers (B) + all have ✓ some have 0 almost none have	To pick out superior workers (S) + very important ✓ valuable 0 does not differentiate	Trouble likely if not considered (T) + much trouble ✓ some trouble 0 safe to ignore	Practical. Demanding this element, we can fill (P) + all openings ✓ some openings 0 also no openings	Columns	S×P	T	Item Index (IT) SP+T	Total Value (TV) IT+S −B−P	P' (+=0 ✓=1 0=2)	SP'	Training Value (TR) S+T+ SP'−B
					9–12							
					13–16							
					17–20							
					21–24							
					25–28							
					29–32							
					33–36							
					37–40							
					41–44							
					45–48							
					49–52							
					53–56							
					57–60							
					61–64							
					65–68							
					69–72							
					73–76							
					77–80							

U.S. Civil Service Commission
Personnel Research and Development Center
Washington, D.C.

Note: for all categories except P', + counts 2, ✓ counts 1, 0 counts 0. For category P', + counts 0, ✓ counts 1, 0 counts 2.

Figure 1. The Job Element Blank

TABLE 1

ELEMENT		BARELY ACCEPTABLE B	SUPERIOR S	TROUBLE T	PRACTICAL P	TV	IT	TR
26	Ability to work under pressure	50	85	67	64	75	59	82
14	Judgment	46	92	84	69	92	69	96
70	Tact & Diplomacy	42	80	61	61	69	53	80
66	Ability to set priorities	38	88	73	65	86	62	92
5	Ability to plan and organize work	45	75	80	80	75	67	70
71	Emotional stability	60	85	65	75	73	65	65
72	Judgment as to when to act	30	90	80	65	93	63	105
1	Ability to gain cooperation	4	94	78	72	97	72	86
28	Ability to handle a no. of tasks at one time	33	89	78	56	89	59	106
104	Initiative & aggressiveness	44	83	67	61	72	56	86
6	Ability to work independently	29	93	86	71	104	71	104
21	Ability to get along w/ higher management & staff	44	83	78	78	81	67	64
29	Ability to shift priorities	50	90	75	70	83	65	88
42	Ability to plan coordinate work	50	78	72	67	67	57	78
27	Ability to meet short deadlines	50	85	65	70	75	62	75
31	Ability to deal problems	36	64	50	57	46	40	68
75	Oral expression	36	100	71	71	104	71	96
11	Ability to deal w/ people outside	43	92	93	79	104	79	93
46	Reliability & dependability	55	90	75	75	85	70	78
69	Judgment inform higher authority	44	88	81	69	88	67	91
103	Analytical ability (troubleshooting)	50	100	75	67	96	70	96
130	Giving workers feedback	50	79	64	64	68	55	79
4	Ability to determine procedures for handling unique problems	40	60	70	80	50	53	40
51	Enthusiasm in work	38	100	50	63	88	58	94

TABLE 1 (cont.)

		BARELY ACCEPTABLE	SUPERIOR	TROUBLE	PRACTICAL			
ELEMENT		B	S	T	P	TV	IT	TR
108	Ability to make on-the-spot decisions	38	63	63	75	50	50	63
121	Reasoning ability	25	100	100	67	121	78	121
123	Ability to speak persuasively	40	80	70	70	75	60	80
50	Keeps information to self	79	57	64	86	21	50	36
62	Ability to impart information	40	100	80	70	105	73	100
112	Personal conduct & integrity	50	75	75	58	67	56	79
120	Ability to handle many jobs at once	67	100	67	50	75	56	100
2	Gain conformance w/out authority	50	63	50	63	44	46	50
63	Interpret written instructions	60	80	80	60	70	60	80
84	Agency operations & procedures	50	67	67	67	58	56	58
109	Knowledge of procedures	60	90	90	80	95	80	75
15	Cooperative w/ others	63	100	88	88	106	88	75
45	Taking on new assignments	25	63	50	63	56	46	63
57	Sense of humor	40	60	40	70	35	40	50
68	Interpret & apply oral instructions	33	100	67	50	100	75	88
106	(N=2) Salesmanship	25	100	50	75	100	75	88
131	Give information to employees	63	63	63	75	38	50	50
132	Get information from employees & acting on it	50	88	63	50	70	50	94
137	Understanding EEO	50	88	63	50	69	50	100

abbreviated form (designated by number from the original job element list), and in approximate rank order according to the number of items each was rated by the entire group of managers (e.g., "ability to work under pressure" was rated by 14 managers and was the single most frequently chosen item).

According to Primoff's (1975) interpretation, the transmuted total value (TV) indicates the extent to which an item covers a broad range of ability, and significant elements will usually have a transmuted TV of 100 or over. Primoff(1975) suggests that in specific situations, it may be advisable to use items with TVs of less than 100 as elements or items with TVs greater than 100 as subelements. Thus, results of the job element analysis may be interpreted with some flexibility. The item index (IT) identifies subelements, and transmuted IT values of 50 or higher indicate significant subelements. The training value (TR) indicates elements or subelements which might be considered in a training program; ability to learn any item having a TR of 75 or more may be considered a subelement. Primoff also suggests that items having high transmuted S values (80 or more) are useful, and items with B values less than 50 should not be eliminated.

As shown in Table 1, seven items have TVs greater than 100, suggesting these are elements. These items and the respective TVs are: (#121) reasoning ability, 121; (#15) cooperative with others, 106; (#62) ability to impart information, 105; (#6) ability to work independently without immediate supervision, 104; (#75) ability to express oneself orally, 104; (#11) ability to deal with people outside the organization, 104; (#68) ability to interpret and apply instructions given orally, 100. In addition, the following items: (#14) judgment; (#72) judgment when to act independently; (#1) ability to gain cooperation; (#103) analytical ability; and (#109) knowledge of procedures, had TVs greater than 90, coupled with high IT or TRs.

In order to define dimensions for the assessment center based on these results, the items were grouped as follows:

Element #121 Reasoning ability

Element #68 Ability to interpret and apply instructioning given orally

Subelements #103 Analytic ability

#4 Ability to handle unique problems

#63 Ability to interpret written instructions and regulations

This group of items appears to comprise a dimension labelled Reasoning and Analytical Ability, which was defined as the ability to draw inferences or conclusions from known or assumed facts to identify existing and potential problems and possible solutions.

A second grouping consisted of the following:

Element #14 Judgment

Element #72 Judgment as to when to act independently

Subelements #31 Ability to deal with problems which are not set in definite terms

#69 Judgment to inform higher authority when necessary

Subelements (cont.)
#108 Ability to make on-the-spot decisions

This dimension was called Judgment and Decision Making, and is defined as appropriately deciding when to act independently and when to inform or refer to higher authority. Ability to evaluate alternatives and choose the most appropriate course of action. Ability to deal with ambiguous problems, make immediate decisions and establish courses of action.

A third group consisted of the following:

Element #15 Cooperative with others

Element #11 Ability to deal with people outside the organization

Subelements #70 Tact and diplomacy

#1 Ability to gain cooperation

#21 Ability to get along with higher management and staff

These items suggested an Interpersonal Skills dimension, defined as the ability to perceive and react to the needs of others and to accurately perceive one's impact on them. Skill in both offering and gaining cooperation from those inside and outside of the organization. Displays tact and diplomacy in all interactions.

A fourth group of two elements and one subelement consisted of:

Element #75 Ability to express oneself orally

Element #62 Ability to impart information

Subelement #131 Giving information to employees

This Communication Skills dimension was defined as the ability to clearly and concisely express ideas in written and oral form in group and personal presentations.

A fifth major group of items was:

Element #6 Ability to work independently without immediate supervision

Subelements #66 Ability to set priorities on work

#5 Ability to plan and organize work

#29 Ability to shift suddenly to new tasks when priorities change

#42 Ability to plan coordination of work of several others, in terms of needs of particular task

#27 Ability to meet short deadlines

#4 Ability to determine procedures (for unique problems)

#120 Ability to handle many assignments at once

#28 Ability to handle a number of tasks at one time

This dimension, labelled Planning and Organization was defined as the ability to independently establish clear cut attainable objectives and foresee resources needed for task accomplishment (men, money and materials). Ability to establish and/or quickly adjust priorities to attain objectives and meet unique problems, changing situations and short deadlines. Ability to effectively coordinate simultaneous assignments for self and others.

The final dimension identified was Leadership and Supervision, and consisted of the following items:

Element #123 Ability to speak persuasively to groups

#2 Ability to gain conformance, without authority

#132 Getting information from employees and acting on it

#104 Initiative and aggressiveness

#130 Giving workers feedback on their performance

This dimension was defined as the ability to guide and direct others and initiate action to influence events. Constructively uses influence and persuasion to gain conformance with or without authority. Ability to motivate, appraise and assist others in job accomplishment and career development.

Definitions of these dimensions were based primarily on results of the job element analysis. While the grouping of elements and subelements was arbitrary, the dimensions so identified were congruent with information obtained in interviews and also with dimensions used in other assessment centers. Definitions were derived from all these sources, with an attempt to link definitions to the element and subelement groups. Thus, this job element approach to identification and definition of dimensions provides an appropriate technique for development of assessment centers. Exercises used in the assessment center were designed to measure the six dimensions identified and defined in this manner.

References

Byham W.C. Assessment Centers for spotting future managers. Harvard Business Review, 1970, 48(4), 150-160.

Flanagan J.C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

Jeswald, T.A. Issues in establishing an assessment center. In J.L. Moses and W.C. Byham (Ed's), Applying the assessment center method. New York: Pergamon Press, 1977.

Primoff, E.S. How to prepare and conduct job element examinations. U.S. Civil Service Commission: Technical Study 75-1, 1975.

Slivinsky, L.W., and Desbiens, B. Managerial job dimensions and job profiles in the Canadian Public Service. Studies in Personnel Psychology, 1970, 2, 36-52.

U.S. Equal Employment Opportunity Commission, Civil Service Commission, Departments of Labor and Justice. Uniform guidelines on employee selection procedures. Federal Register, Vol. 43, No. 166, Aug., 1978.

APPENDIX A

Calculating values for the four response categories requires a summation of response values for the items, with the meaning of the summary scores implicit in their definitions; e.g., a high B sum indicates that most barely acceptable workers are satisfactory, etc. However, since the purpose of the calculations is to find elements which will select superior workers, summary scores must be modified by other considerations. For example, a high score on "superior" must be modified by the consideration as to whether it is practical to expect the element; this modification is obtained by multiplying S x P, and a weight is added to indicate trouble if the element is ignored (SP + T), resulting in a value called the Item Index (IT).

Since the numerical size of each group sum is affected by the number of raters, it is necessary to transmute group sums in order to provide a scale of values which is constant across all groups of raters. The transmuted Item Index is calculated by dividing the group IT sum (applying the formula SP + T) by a base of six (since six is the maximum possible individual rating) times the number of raters, and multiplying the result by 100.

The Total Value (TV) of an item indicates whether it is broad, i.e., whether the difference in ability between barely acceptable and superior workers is great, or is relatively narrow and considered to be a subelement. The formula for computing TV is IT + S - B - P. High TV items are considered to be major elements. Transmuting group values requires calculating the TV formula using group sums, dividing this result by four times the number of raters, and multiplying by 100.

Similarly, a training value (TR) is obtained by the formula S + T + SP' - B (P' is obtained by reversing the values assigned to zero and plus responses on P, i.e., zero is scored two, and plus becomes zero). The transmuted TR value is obtained by applying this formula to group sums, dividing the result by four times the number of raters, and multiplying by 100.

Detailed information regarding the rationale, scoring procedures, and interpretation of results is given by Primoff (1975).

ASSESSMENT CENTER PREDICTIONS OF THE OFFICER EVALUATION REPORT

Frederick N. Dyer and Richard E. Hilligoss

US Army Research Institute Field Unit
Fort Benning, Georgia 31905

INTRODUCTION

In 1973-1974 the US Army Infantry School (USAIS) Assessment Center (ACTR) assessed students from the Infantry Officer Advanced Course (IOAC), the Infantry Officer Basic Course (IOBC) and the Advanced NCO Educational System (ANCOES) to determine the feasibility of the assessment center as a technique for leadership development and leadership prediction. Students from the Branch Immaterial Officer Candidate Course (BIOCC) were assessed also to determine the feasibility of the assessment center as a selection device.[1]

Dyer and Hilligoss[2] related the ACTR scores on these officers and NCOs to ratings of field leadership obtained six months following completion of leadership training and assignment to new duty stations. These ratings were mady by supervisors, peers and subordinates of the former assessees on a fifty item questionnaire, "Leadership Performance Rating Form" (LPRF). A second study[3] related the ACTR scores to predicting a second criterion, the end-of-course grade obtained by the assessee in the leadership course completed immediately after going through the assessment center. The purpose of the present paper is to discuss the effectiveness of the ACTR for predicting a third criterion, the Officer Evaluation Report (OER). Since the OER is applicable to officers only, the discussion will be confined to the three Officer assessment groups.

METHOD

ACTR Description

Table 1 presents a summary of assessee characteristics and group sizes.

[1] US Army Infantry School Assessment Center After Action Report: Executive Summary (Book 1, Volume 1), December 1974, DDC Number ADB001183L.

[2] Dyer, F. N. and Hilligoss, R. E., Using an Assessment Center to Predict Field Leadership Performance of Army Officers and NCOs. Proceedings of the 19th Annual Conference of the Military Testing Association, October 1977.

[3] Dyer, F. N. and Hilligoss, R. E., Using an Assessment Center to Predict Leadership Course Performance of Officers and NCOs. Proceedings of the 20th Annual Conference of the Military Testing Association, October 1978.

Table 1

ASSESSEE GROUP CHARACTERISTICS AND SIZES

	ASSESSMENT GROUP		
Descriptor	IOBC	IOAC	BIOCC(OCS)
Number Assessed	90	88	143
Pay Grade	O-1	O-3	E 3-6
Average Age	22.6	28.8	25.3
Average Years of Active Duty	0.3	5.7	3.3
Number with OER ratings	69	67	84
Number with complete LPRF Data	45	36	40
Number completing leadership courses	87	84	105

Assessment Center Personnel

The assessor pool consisted of six Majors, seven Captains, two Lieutenants, three Master Sergeants, two Sergeants First Class, and one Staff Sergeant. Assessors were selected by DA using the following criteria: each man must be in one of the Combat Arms; each Captain and above must have had command experience; each Major, Captain and Sergeant must have served in combat, and officers must have an advanced degree in one of the behavioral sciences. The assessors received training on principles and techniques in assessment, interviewing and counseling for four months before beginning their duties. The training included repeated rehearsals of assessment exercises.

Behavioral Dimensions

The twelve dimensions of leader behavior that were judged to be appropriate for measurement in the Assessment Center were:

- Adaptability
- Administrative skills
- Communication skills
- Decision Making skills
- Forcefulness
- Mental Ability
- Motivation
- Organizational leadership
- Physical fitness
- Social skills
- Supervisory skills
- Technical and tactical competence

Assessment Center Exercises

Group Exercises:
- Leaderless Group Discussion
- Management Exercise (Conglomerate)
- Rotating Assigned Leadership Exercise (Field exercise)
- Leader War Game (IOAC)

Interpersonal interaction (one-on-one):
- Entry interview
- Appraisal interview

Work samples:
- In-basket
- Writing exercise

Leadership in a simulated situation:
- Radio simulate

Psychometric Tests

- Henmon-Nelson Test of Mental Ability
- Nelson-Denny Reading Test
- Social Insight Test (Chapin)
- Strong Vocational Interest Blank
- Watson-Glaser Critical Thinking Appraisal

Self-description Instruments

Edwards Personal Preference Schedule
Leadership Opinion Questionnaire (Fleishman)
Leadership Q Sort (Cassel)
Person Description Blank
Work Environment Preference Schedule (Gordon)

THE US ARMY OFFICER EVALUATION REPORT AS A LEADERSHIP CRITERION

The OER (DA 67-7) consists of ten parts of which parts V and VI provide the numerical score. Part V, "Demonstrated Performance of Present Duty" consists of six descriptors from "inadequate" to "outstanding" with a maximum total of 70 points. Part VI, "Potential" consists of five descriptors from "I would not promote this officer," to "Promote this officer immediately" with a maximum total of 30 points. A maximum score of 200 is possible by combining top scores for the rater and the indorser.

An officer who is "effective" and who "should be promoted with his contemporaries" could theoretically receive an OER rating as low as 46. However, almost all officers are rated within a few points of the 200 maximum and higher grades of officers have higher average ratings. Despite this partial "ceiling" effect, sufficient variance existed in the OERs for the assessees to produce correlations between the most recent OER which was used as the criterion and four previous OERs which averaged .50 for the IOBC assessees, .47 for the IOAC assessees and .62 for the BIOCC assessees. These correlations indicate substantial reliability of these measures. The high relevance of the OERs for Officer advancement makes them an important criterion for validation of the USAIS Assessment Center.

RESULTS

Table 2 lists the percentage of successful ACTR Predictors ($p<.05$) of the three leadership criteria for the different assessment groups. The figure "8.82" percent, for example, is obtained by dividing the six successful predictors by the maximum number of predictors, in this case sixty-eight. In a similar manner, the rest of the percentages were obtained.

Those exercises having assessor ratings were Leaderless Group Discussion, Conglomerate, Radio Simulate, In-basket, Appraisal Interview, Writing Exercise, Assigned Leader Group Exercise, and Leader Game (IOAC). Those exercises having peer and self-ratings were Leaderless Group Discussion, Conglomerate, Assigned Leader Group Exercise and Leader Game. Self-Description instruments included Edwards Personality Preference Schedule, Work Environment Preference Schedule, Leadership Opinion Questionnaire, Leadership Q Sort and Person Description Blank.

Considering the OER as the criterion under Assessor Ratings Formal Exercises, "Presentation impact," ($r=.25$) on the Leaderless Group Discussion for the Army Lieutenant was related to the criterion. "Forcefulness" ($r=.31$) and "effectiveness in an organizational leadership role," ($r=.33$) on the Radio Simulate for

TABLE 2

PERCENTAGE OF SUCCESSFUL ACTR PREDICTORS (p.<05) OF
THREE LEADERSHIP CRITERIA FOR DIFFERENT ASSESSMENT GROUPS

Class of ACTR Score	No. of Scores Per Assessee	OFFICER EVALUATION REPORT			LEADERSHIP PERFORMANCE RATING FORM			END-OF-COURSE GRADE		
		IOAC	IOBC	BIOCC	IOAC	IOBC	BIOCC	IOAC	IOBC	BIOCC
Assessor Ratings Formal Exercises	68	8.82	8.82	22.06	5.88	7.35	5.88	51.47	45.59	60.29
Peer Rankings Formal Exercises	15*	0	6.67	33.33	0	0	6.67	81.25	53.33	60.00
Self-Rankings Formal Exercises	15*	0	0	26.67	6.25	0	0	81.25	6.67	0
Entry Interview	14	7.14	0	14.29	0	7.14	42.86	28.57	50.00	50.00
Pencil & Paper Performance Tests	9	0	66.67**	33.33	0	0	0	100.00	55.56	88.89
Self-Description Instruments	75	12.00	26.67	4.00	20.00	13.33	5.33	25.33	10.67	16.00

*16 for IOAC

**All significant correlations were negative.

the BIOCC assessees were related to the criterion. For the Army Captain, "supervision," (r=.30), and "directing skill," (r=.25) on the In-basket were related to the criterion. The OER rating on the Army BIOCC candidate was related to "oral communication," (r=.31), "sociability," (r=.37) and "overall effectiveness," (r=.28) as rated by peers on the Management Conglomerate. "Quantitative," (r=.27) on the Henmon-Nelson Test of Mental Ability was related to the criterion for the BIOCC candidate. "Strong," (r=.26) on the Person Description Blank was related to the criterion for the IOAC Captain.

By far, the BIOCC group accounted for the largest number of successful predictors. Table 3 gives the contribution of each separate exercise having an assessor rating to the total.

Considering the Leadership Performance Rating Form (LPRF) and the OER as criteria an explanation of the failure of the traditional assessment center exercises to predict the field leadership ratings and the OER was that something other than leadership was being rated by the superiors, peers and subordinates (LPRF) and the rater and endorser (OER) who provided these ratings. The success of the self-description instruments in predicting these two criteria (relative to assessor ratings) suggests that little opportunity had existed in these peacetime field settings for leadership to emerge and, in its absence, the leader's self-perception was communicated to the raters and used as the basis for leadership ratings.

Table 3

PERCENTAGE OF SUCCESSFUL FORMAL EXERCISE PREDICTORS ($p<.05$)
FOR THE BIOCC GROUP WITH OER AS THE CRITERION

Exercise	Assessor Ratings	Peer Rankings	Self-rankings
Writing Exercise	50	NA	NA
Radio Simulate	38	NA	NA
In-basket	29	NA	NA
Management Exercise	12	20	40
Appraisal Interview	12	NA	NA
Leaderless Group Discussion	11	67	17
Assigned Leader Exercise	0	25	25

Table 4 gives the intercorrelations between the three leadership criteria. Column 1 shows that there is a "moderate" correlation between the Last OER and the LPRF. These correlations between the Last OER and LPRF further suggest that whether rating by the OER format or rating by the LPRF format, the rater conceptualizes leadership in a similar manner. Therefore it is not surprising that few successful predictors existed for the IOBC and IOAC groups when OER was the criterion. Evidently, what raters conceptualize as leadership for both OER and LPRF are different from ACTR raters.

An explanation of the higher precentage of ACTR predictors of the OER for the BIOCC group is that these 2d Lieutenants, as platoon leaders, were observed more often and in greater depth in traditional leadership positions and their raters tended to rate more like the ACTR raters.

For the end-of-course criterion, the paper and pencil tests provided the largest proportion of criterion predictors, followed by Formal exercises. It is not surprising that the paper and pencil tests provided the largest proportion of criterion predictors since an end-of-course academic grade reflects, in part, the student's reading and comprehension skills; factors which weighted heavily in the paper and pencil test scores. What is of considerable interest is that the traditional staples of Assessment Centers, i.e., the assessor rating on formal exercises predicted this course grade criterion so well.

Taken overall, the results indicate only marginal utility of the USAIS ACTR for prediction of the OER for company grade and below officers. However, an end-of-course questionnaire indicated a belief by the assessee that the Assessment Center served to make him more aware of his leadership strengths and weaknesses by forcing him to discuss them when video-tape playbacks were reviewed as part of the counseling session. Many have said, "Why hasn't the Army done this for me before?" and "The best thing the Army has done."

Table 4

INTERCORRELATIONS BETWEEN THE THREE LEADERSHIP CRITERIA

Assessment Group	Last OER with Average LPRF	Last OER with End-of-course Grade	Average LPRF with End-of-Course Grade
IOAC	.39	.13	.30
IOBC	.33	.04	.15
BIOCC	.45	.43	.25

Largely because of assessee reactions and leadership course grade predictions the Army is going ahead with an assessment type program for ROTC cadets. Under contract, it is hoped, in connection with this new research, that the best of the ACTR exercises can be automated to reduce assessment cost.

A ten year follow-up of the USAIS ACTR will use promotion of the Infantry Officer and NCO as the leadership criteria.

A FURTHER EVALUATION OF ROTC GRADUATE PERFORMANCE

by

Arthur C. F. Gilbert, Ph.D.

A Paper Prepared for Presentation at the
21st Annual Conference of the Military Testing Association (MTA)
San Diego, California
October 15 - 19, 1979

Performance and Training Research Laboratory
U. S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

A FURTHER EVALUATION OF ROTC GRADUATE PERFORMANCE

Arthur C. F. Gilbert, Ph.D.
U. S. Army Research Institute
for the Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

Previous research (Gilbert, Weldon, and Wellins, 1978) dealt with the evaluation of a sample of Army Reserve Officer Training Corps (ROTC) graduates in Officer Basic Courses (OBC) and evaluated their performance in relation to officers entering on active duty through other officer procurement programs. This research also compared the OBC performance of ROTC graduates who were scholarship recipients with the performance of ROTC graduates who were not ROTC scholarship recipients. In addition, the OBC performance of male ROTC graduates was compared with the performance of female ROTC graduates, and the performance of ROTC graduates who attended ROTC institutions in the four ROTC regions was contrasted. This initial research reported the findings for a sample of Army officer accessions in 1977. The authors concluded that "the Army ROTC program is producing a quality of graduate whose performance in the Officer Basic Course is of a comparable quality with that of other officer procurement programs."

The purpose of this research was to determine if the findings of the earlier research could be replicated for another sample of officer accessions. Specific objectives as in the earlier research were to (1) compare the performance of ROTC graduates with that of officers entering on active duty from the other officer procurement programs; (2) compare the performance of ROTC scholarship recipients and ROTC graduates who were not scholarship recipients; (3) compare the performance of male ROTC graduates with the performance of female ROTC graduates [illegible] (4) compare the performance of ROTC graduates who graduated from colleges or universities located in the four ROTC Regions. The criterion of performance was the final course grade obtained in the OBC of the Army Career Branches.

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

PROCEDURE

A sample of 1,561 officers who completed Officer Basic Courses (OBC) in 1978 was used as subjects in the present investigation. As in the previous investigation, the sample was divided on the source of commission. However, in this investigation only three groups representing three of the sources of commission were used because there was not sufficient representation in the sample of officers receiving direct appointments. Therefore, the three groups of officers according to source of commission were: ROTC, U. S. Military Academy, and Officer Candidate School (OCS). The ROTC sample was divided on the basis of ROTC scholarship status, ROTC geographical region corresponding to the location of the ROTC host institution from which they graduated, and on the basis of sex.

The criterion used in the analyses was final course grades earned in the OBC. OBC final course grades were converted to Army Standard Scores within the different OBC. Analysis of variance was used to determine the significance of the differences in means of the groups in the four separate analyses.

RESULTS AND DISCUSSION

In Table 1, the means of the three different procurement programs are shown for the analysis of variance for the 1978 sample. The means of OBC final course grades for ROTC scholarship recipients and non-recipients is also presented as well as the mean performance of male and female ROTC graduates and the mean performance of ROTC graduates from the four geographical ROTC regions.

The results of the analysis of variance indicated a significant difference among the means of the three groups of officers. Graduates of the USMA were favored while ROTC graduates were next and the lowest mean OBC course grade was for graduates of the OCS program. Comparison of the mean performance of USMA and ROTC scholarship recipients indicated that ROTC scholarship recipients were slightly favored in terms of OBC final course grade average but that this difference was not significant.

ROTC scholarship recipients performed significantly better than non-scholarship recipients; the mean Army Standard Score for ROTC scholarship recipients was 108.06 compared with the mean of 95.76 for non-scholarship recipients. Male ROTC graduates were slightly favored over female ROTC graduates; the mean OBC final course grades being 108.48 and 98.30 respectively, but the difference in means between the two groups was not statistically significant. The mean performance of ROTC graduates among the four ROTC regions yielded a significant difference. In decreasing order of magnitude the means for the four ROTC regions were Western ROTC Region, 105.82; North Central ROTC Region, 102.89; Eastern ROTC Region, 98.42; South Central ROTC Region, 96.69.

TABLE 1

MEAN OFFICER BASIC COURSE FINAL GRADES
FOR THE DIFFERENT GROUPS IN THE
1977 SAMPLE OF OFFICER ACCESSIONS

Group	N*	$\bar{X}$
U. S. Military Academy	664	107.31
ROTC	466	99.97
OCS	69	98.34
Total	1,379	102.27
ROTC Scholarship Recipients	152	108.06
Non-Recipients	314	95.76
Male ROTC Graduates	1,327	108.48
Female ROTC Graduates	234	98.30
Eastern ROTC Region	540	98.42
North Central ROTC Region	243	102.89
South Central ROTC Region	267	96.69
Western ROTC Region	200	105.82

*All 1,561 cases were not used because of missing data elements.

TABLE 2

COMPARISON OF MEAN OFFICER BASIC COURSE (OBC)
FINAL GRADES OF THE 1977 SAMPLE
AND OF 1978 SAMPLE OF OFFICER ACCESSIONS

	1977 Sample	1978 Sample
U. S. Military Academy	106.90	107.31
ROTC	100.34	99.97
OCS	96.22	98.34
Total	100.20	102.27
ROTC Scholarship Recipients	105.81	108.06
Non-Recipients	96.72	95.76
Male ROTC Graduates	100.48	108.48
Female ROTC Graduates	98.30	98.30
Eastern ROTC Region	98.13	98.42
North Central ROTC Region	101.99	102.89
South Central ROTC Region	96.83	96.69
Western ROTC Region	106.14	105.82

In Table 2, the means of the groups for the different procurement programs are shown for the 1977 sample as well as the means of the different ROTC groups of graduates for that research. Also, the means for the current research are shown for comparative purposes. In the earlier research, significant differences among the four procurement programs were obtained with the mean of USMA graduates being favored. A significant difference between USMA graduates and ROTC scholarship recipients was not obtained. As in the present research, ROTC scholarship recipients had a higher mean performance than non-recipients. Male ROTC graduates were slightly favored over female ROTC graduates but there was not a statistically significant difference between the means for the two groups. The differences among the four ROTC regions were also significant with the means for the four regions being in the same relative order of magnitude.

Generally, the results of this research support the results of the earlier reported research (Gilbert, Weldon, and Wellins, 1978). The Army ROTC program continues to produce officers whose performance in OBC is comparable to that of officers from the other commissioning programs. Future research will focus on isolating the critical dimensions of officer performance during initial tours of active duty within the different officer specialities. Rating scales will be developed to measure performance along these dimensions to isolate more specific differences in performance and provide more meaningful information than the single index of performance used in the present research.

REFERENCE

Gilbert, A. C. F., Weldon, J. I., Jr., & Wellins, R. S. Quality of ROTC ac sions to the Army officer corps. Paper presented at the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK, October 30-November 3, 1978. <u>In Proceedings, 20th Annual Conference of the Military Testing Association.</u> Oklahoma City, OK: U. S. Coast Guard Institute, 1978.

PSYCHOLOGICAL MANAGEMENT TRAINING IN THE BUNDESWEHR

-A model endeavour to exercise senior officers in the practice of social skills-

Klaus J. PUZICHA

Federal Armed Forces' Psychological Institute
Dezernat Wehrpsychologie im Streitkräfteamt

Summary

The command and control circumstances surrounding military superiors at the level of battalion commander in the Bundeswehr has been marked by the following developments:

1. The steadily increasing degree of democracy, technology, and work sharing.

2. The increasing complexity and specificity of control and command tasks.

3. The increasing heterogeneousness of the men in respect of their qualifications, interests, and needs.

Proceeding from an empirical situation analysis (document analysis, observation of activities, and interview of experts) a general curriculum has been elaborated to train future commanders at the said level. Within the framework of the attainment targets concerned, viz

- the application of a knowledge of socialpsychology in the prevention and resolution of social conflict and the promotion of motivation and

- the shaping of planning and deciding processes in the light of the technical competence and needs of the men,

the Federal Armed Forces Psychological Service has elaborated and tested a programme of training marked by

1. the principle of encouraging trainees to play an active part (playacting, group work, and group dynamic and communication exercises),

2. the principle of problem-oriented work so as to ensure that what has been learnt can be extrapolated to other fields, and

3. the principle that method takes priority over factual knowledge.

1. Preliminary Remarks

Due to the situation peculiar to German rearmament in the mid-fifties particular importance attached to the intellectual foundation of the Bundeswehr right from the beginning. The questions arising from fitting the forces into the structure of the Constitution were parallelled by the problem of interpolating them into the system of values of the liberal democratic basic order of the Federal Republic of Germany, still only a few years old. Discussion on these matters eventually found expression in the conception of General BAUDISSIN entitled "Innere Führung" (Leadership and Civic Education), which had two functions, each at a different level and each complementary to the other. They were

- an integrational function intended to ensure both the social and the political integration of the forces - in other words to prevent their developing into a state within the State,

  and

- a motivational function intended to ensure that the individual serviceman, wether regular or conscript, was in a position to see his service, "from his daily experience of freedom, human dignity, and political reasoning" (HESSLEIN, 1977, p. 7), as a purposeful and important mission and act accordingly.

The relative importance of these two functions - dependent upon military and social developments during the past twenty years - has undergone several shifts. In the beginning - if only due to the distrustful watchfulness with which foreign eyes followed German rearmament - it was integration which came to the fore; now it ist motivation. Problems arising from unsuccessful motivation make their appearance mainly in conscripted men in multivarious forms of trend towards deviant behaviour and motivational deficits; however, they also appear in the form of behavioural insecurity of superior soldiers, in their consequent search for new patterns of interpretationand maxims of behavior, which can sometimes lead to retroalignment with traditionalist patterns of values and ways of acting. (Cf. GESSENHARTER et alia, 1978).

Such a -trendwise - insecurity of military superiors is very probably determined by the following developments also:

- the growing measure of democracy, technology, and work sharing, in the daily life of the forces,

- an increasing measure of complexity and speciality in command and control task, and

- a rising measure of heterogeneousness in the qualifications, interests, and needs, of subordinates (cf. SEITZ et alia, 1978, p. 73).

Command behaviour training in the Bundeswehr to day has mainly been in the form of on-the-job training; thus far, there has never been any systematic behavioural training directed towards flexible, cooperative command behaviour adequate to dealing with the situation an hand. Proceeding from this state of affairs, a programme of training has been elaborated for officers holding commands at battalion level which - as a model - contains a programme of psychological training for behaviour modification and sensibilisation. The following is an account of this programme.

2. The Elaboration of a Curriculum of Training for Future Air Commanders

2.1 Target Group

"As a rule, trainees are regular officers, major or lieutenant colonels, aged between 35 and 41. About 80 percent are qualified for entry into a university, the remainder for entry into a technical college. The present exceptions to the rule are trainees trained in science (this will change in ten years or slightly over due to the training given to officers at Bundeswehr universities). Trainees' military careers have usually taken them through a unit command and staff work.

Trainees already exercising the functions of a commander or second in command are exceptions.

A course numbers up to 15 trainees. They come from every branch of the Air Force, where they have been concerned mainly "on the practical side" with daily military problems of limited complexity. They have an independent style of working. Although they have had experience of courses during their military training, their "reteurn to school" calls for considerable readjustment in most cases. Since the men on such courses will be assuming the duties of a commander or second in command immediately after the course or within a short time, there is always sufficient motivation for voluntary cooperation while on the course. They expect, for instance, that such a practical course will provide them with the aids and tools requisite for their future assigments, albeit they are very critical of any training which fails to come up to such expectations." (SEITZ et al., 1978, p. 77 et seq.)

The motivation to learn - it may safely be said - is more of an intrinsic nature, since the courses are completely free from sanctions. There are no assessments and no marks are awarded; there is no question of passing or failing.

2.2 Analysis of the Professional Situation of an Air Force Commander

Before a curriculum for training future Air Force commanders was elaborated, an attempt was made "to ascertain as comprehensively and extensively as possible the situation of such a commander. This step was termed a situation analysis." (SCHWANS et al., 1978, p. 11 et seq.)

Three complementary procedures were employed to collect data, viz a document analysis, participating observation of commanders'activities, and interviews with experts.

The Document analysis included "all the regulations, orders, instructions, and ordinances, relevant to the general military field of activity of a commander" (SCHWANS et al. in passage loc.cit., p.12). The analysis provided a total of roughly 650 items of information on commanders' activities.

The participating observation of activities classified the daily routine of a commander in accordance with the following work-analytical questions:

- How is a problem solved?
- What is the object of the performance involved?
- Wat means were employed?
- When is the task accomplished?
- Where is the task accomplished?

and recorded the answers in an summary data sheet. The observation covered roughly 700 hours, and was complemented with interviews.

In the interview of the experts all the present commanders, their immediate superiors, and a sample of their immediate subordinates, were questioned upon

- their estimation of the practical and professional tasks involved,
- their estimation of the command and control functions involved, and
- their views on the course given to future commanders.

The questions were asked by correspondence; the response rate was 85 %.

2.3 Selected Results

The findings of the empirical situation analysis were (cf. SCHWANS et al., 1978, p. 29):

1. The work of a commander is largely determined by tasks dictated by Innere Führung, personnel management and records, welfare, public relations work, training, organisation, and control and command.

2. His main activity ist "staff work".

3. A commander is insufficiently trained for his job.

4. Most of his problems occur in the field of "Training and Organisation".

5. The main reasons for there being problems are complexity and lack of clarity in the tasks to be accomplished.

6. The present course for future commanders in the Air Force offers too little in the way of aids to practice.

7. The subjects of "military command and control procedure and staff work" must be handled with priority in any commanders' course.

8. Commanders' courses should in future be based mainly on practical cases.

9. From the point of view of method, conversational forms of training, especially in the shape of plays, are awared a high priority.

The curriculum is elaborated in five separate steps. Following a description of a commander's situation in respect of funktion and training (summarised above) a list of learning objectives was worked out. This was followed by the "organisation of learning objectives", which is to say they were broken down by subjects and the time available. Each learning objective in the overall curriculum was subdivided into training objectives in a syllabus.

The overall curriculum was designed in the light of three determinative criteria, viz

- flexible curricular procedure, i.e. although it "contained a systematically elaborated catalogue of learning objectives, it was sufficiently flexible to leave teacher and trainees at the baseline enough room for decision and action to take into consideration their own interests and needs." (SCHWANS et al., loc.cit., p. 51.)

- A curricular procedure in direct relation to the situation, which is to say an approach dictated by an overall view, in accordance with practice, of the field of activities, and not by the traditional subject-by-subject system used by training personnel.

-A curricular procedure aligned with the learning objectives i.e. "aimed at the desired qualifications, which should be described as clearly and unambiguously as possible, which is to say fields of knowledge, abilities, skills, and attitudes." (SCHWANS et al. loc.cit., p. 52.)

From a total of 41 learning objectives we select two as examples to elucidate a military psychological programme of training which - as a model - is to provide a systematic behavioural training to exercise social skills. The learning objectives in question are

1. the use of socio-psychological knowledge in preventing and resolving social conflicts and promoting motivation and

2. the shaping of planning and decision processes in the light of the objective competence and the needs of the men.

3. A Military Psychological Programme of Training

The programme of training elaborated by the Military Psychology Service to achieve the learning objectives referred to has now been tested four times by military psychologists in the form of courses for future commanders. The courses included an entire age group of future Air Force commanders. The substance and didactics of each course were modified in the light of experience gleaned from its predecessor, which is why the presentation of the programme following two courses (EBENRETT et al., 1978) is not up to date. However, it is precisely by reason of the continual process of modification on the basis of concrete experience and - in no smaller measure - the interest and need structure of trainees that the programme described below has been developed up to the point at which it can be recommended for extension to the entire Bundeswehr, which is to say to army and naval officers at comparable level.

3.1 Training Objectives

The Innere Führung (Leadership and Civic Education) code of standards demands of military superiors in the Bundeswehr cooperative (participative) and flexible (i.e. situation-consonant) behaviour as commanders. Such behaviour is marked by the following characteristics (SCHWANS et al. loc.cit., p. 73 et seq.):

- "A commander seeks to legitimise his leading role rather by technical and personal authority than by official authority."
- "He sees his function mainly in relation to its purpose."
- "He enables the men under his command - provided that their qualifications and the situation permit - to participate in command functions."

- "By so doing and by explanatory information he makes his decisions as a commander transparent."
- "Control is not a sign that one ministrusts one's subordinates, but a function executed in the interests of the mission."

On the basis of this catalogue of standards and the two learning objectives referred to above we have defined the following training objectives:

+ Subject: Communication

  - Getting acquainted with a few aspects of group-dynamic objectives and methods.
  - Learning how one's own behaviour affects other people.
  - Learning to express one's own perceptions, ideas, and wishes, more frankly.
  - Acquiring knowledge for an adequate feedback.

+ Subject: Behaviour towards Difficult Men

  - Acquiring information on manifestation, causes, and the possibilities of prevention, in cases of suicidal attempts of soldiers.
  - Acquiring knowledge conducive to psychologically adequate behaviour towards "difficult" men.

+ Subject: General Questions of Military Psychology

  - Getting acquainted with the results of present-day military psychological studies (the problems of conscientious objection, AWOL/desertion, motivation to join up, leisure time behaviour in the forces, analysis of the effects of political education).

3.2. Organisation and Didactic Principles

The following fundamental structures in respect of organisation and didactics are the hallmarks of the training programme:

A. The principle of encouraging trainees to play an active part: small groups (no more than 15), team teaching by three psychologists, changing the learning conditions (individual work, work in small groups, plenary events; schoolroom teaching less than 10 % of the total time) were the methods employed to achieve this end.

B. The principle of centring on the problen; i.e. transferability of what has been learnt to other fields: concrete problematical areas were treated very intensivily - through the medium of psychological sensibilization and the acquisition of a knowledge of psychological teaching - to acquire various new problem situations calling for adequate behavioural techniques.

C. The principle of priority of method over factual knowledge: it is not the acqusition of cognitive information which is the first consideration, but the systematic training of behavioural techniques which lead to solving problems amongst one's fellow men in a manner in keeping with the situation obtaining.
This concentration on modifying and sensitising behaviour at the expense of the communication of cognitive information was attempted through the medium of group-dynamic exercises in interaction and communication and/or by giving the members of the group parts in a roleplay, using the video technique. Cognitive information was imparted mainly by way of written abstracts of, for instance, current studies in military psychology.

The time scale available for these topical learning objectives was seven double periods, which were taken in compact form on two consecutive days.

3.3 Substance

The training objectives referred to above in respect of "behaviour towards difficult men" and "communication" were the crucial points, in roughly equal proportions, in the entire programme. It was mainly left to the trainees to read up the subject of "general questions of military psychology" for themselves from prepared papers.

Psychologically correct behaviour towards difficult men was exercised around the example of men with suicidal attempts. Whilst the suicide figure for the Bundeswehr has been relatively constant over the past ten years, the number of attempted suicides in recent years - by comparis with the figure for non-military groups (students, for instance) -has been rising steadily, so that it is becoming more and more a problem for superior officers.

Behavioural training in this field began with case operation. The trainees were set a training task entitled "attempted suicide by a serviceman", upon which they worked in small groups, elaborating steps to deal with a concrete case presented to them through the medium of an "assortment" of written reports, comments, and additional explanations. The results arrived at in the small groups were subsequently deliberated in a plenary session, and a "psychologically correct" exemplary solution was elaborated. Proceeding from this formal solution, the item of "the commander's talk with the person who attempted suicide" was selected and played - again in small groups -, the members taking the respective parts; the whole was video-recorded. The part of the attempted suicides was played by psychologists so as to ensure systematic variation in behaviour (e.g. in the dimension of depression versus aggression).

Following the play, the results arrived at by the small groups were shown on the screen in a plenary session and discussed. In such deliberations, the following points of view were stressed: locality and organisation of the conversation, choice of time, opening remarks, consistency of non-verbal and verbal communication, bridging gaps in the conversation, specific subjects discussed.

In the subjekt matter of communication a motivation phase taken from the group dynamic exercises was interpolated. In team teaching, talks were given upon the significance and objectives of group dynamik exercises, the relative priority of non-verbal communication in social relation, and the fields of application for experience gleaned in group dynamics in the trainees' professional situation. The group dynamic part was in the nature of a structured, action-centred encounter (cf. BÖDIKER / LANGE, 1975, p. 81). The exercise was conducted in a plenary session with two of the teaching staff. The interaction exercises may be summarised under the headings of "non-verbal behaviour", "trust and frankness", and "clarifying interpersonal relations", the usual feedback rules (e.g. KIRSTEN/ MÜLLER-SCHWARZ, 1976, p. 135 et seq.) became an obligatory code of standards. The types and sequence of the exercises were not scheduled, but proposed by the teachers according to their assessment of the prevailing group situation.

One of the group dynamic meetings, which lasted for about six hours, proceeded roughly as follows:

- An exercise in the "non-verbal search for a partner", in which the trainees were given five minutes to try, using only visual contact as a mean of communication, to come to an unterstanding with a partner that they were desirous of carrying out the next (two-man) exercise with him.

- An introduction exercise in the form of "self-introduction": in groups of two, each of the two had five minutes to introduce himself. The contents of presentation were absolutely free; dialogue, and questions were not permitted. In the plenary session, the officer to whom he had introduced himself introduced him, using the first person and laying his hand upon his partner's shoulder by way of identification.

- Metacommunication phase: how did the trainees feel during this exercise? Why did they feel like that?

- An exercise in self-description: each candidate wrote down on a sheet of paper, anonymously, the three attributes he considered the most salient in his own personality. Each member of the group took one of the sheets and read aloud what had been written on it, upon which a discussion upon who might have been the author ensued. None of the candidates was under any obligation to reveal his identity.

- Metacommunication phase.
- The "hot seat" exercise, which is to say a reciprocal, systematic feedback on the part of the members of the group.
- Metacommunication phase.

3.4 Experience gained

To date, it has not been feasible to conduct a check upon what trainees have learned through the medium of changes in their professional behaviour. The estimation of the success of their training is based upon discussions with all the trainees at the end of the course and questions asked after about six weeks upon the relevance of what had been taught and the measure of consonance between the organisation of the course and the didactics employed on the one hand and the learning objectives on the other.

The motivation of the trainees to learn can be estimated as very strong, and the atomosphere in which the courses were held was found by all concerned to be very good. The great majority considered the basic didactic structures (team teaching, learning by example, changing groups very little schoolroom teaching, behavioural training) to be adequate for the objectives to be attained. Trainees' judgement of the play-acting part of the training was unanimously favourable, whilst their estimation of group dynamic training diverged: well-nigh boundless enthusiasm on the one hand and clearly emotional opposition on the other just about struck a balance. In our estimation the group dynamic exercises provoked in nearly all the trainees a measure of embarrassment and perplexity which very probably paved the way to the behavioural modification envisaged by the learning objectives.

ON THE PROBLEMATIC NATURE OF THE PLANNING ACCORDING TO THE REQUIREMENTS AND THE DEVELOPMENT ACCORDING TO ABILITY OF THE OFFICERS'S CAREER: ANALYSIS OF PROBLEMS AND POSSIBLE SOLUTICNS.

Col. H.E. Seuberlich
Chairman Army Section DBwV (Federal Armed Forces Association).
Südstraße 123, 5300 Bonn 2, W. Germany.

1. Problem analysis

The general outline of the serviceman's profession shows specific characteristics which appear individually in other professions too; nowhere, however, to such an extent. For the regular officer, who is under discussion here, they are particularly applicable.

The point of departure for the following considerations is the ascertainment that the profession of a regular officer must be planned as a "career profession". By this is meant at the present time: the "professional goal" of an officer who begins as a second lieutenant is the rank of lieutenant colonel, for a staff officer tne rank of colonel. Promotions above and beyond this are, however, often hoped for, but they are not a part of the"frame of expectations" which can be attained with certainty by "work performarl e".In addition other factors come into play which cannot be entered into in detail here. Apart from this, the establishment of the officer's profession as a "career profession' is indispensible. Whoever does not want to see it in this light would be unsuited to this profession, at the very least from a mere typological point of view. Several characteristics of central importance which differentiate this profession from others are put forward here:

-the work performance demanded of regular officers in peace time correspondsonly in part with those requirements he must be able to fulfill in wartime.

- Independent of the "simulation" of particular strains and stresses which lead to above average physical and psychical demands in peace time as well, the establishement of "age limits" is imperative as a basis for reassignment possibilities in certain command positions.

- the necessity of having to provide qualified leadership manpower for high and the highest functions too, at short notice for losses in wartime which can be high and can occur very quickly, makes high-grade training courses indispensible for a category of persons which quantatively far exceeds the "precise" needs in peace time.

The features of personnel structure for regular officers specified above are characteristic of the multidimensional area of tension latently surrounding the entire personnel structure of the armed forces, as these features have a continuous bearing upon: supply requirements, ascertainment of supply, planning and measures for the supply of needs and for the development of necessary personnel.

Three factors which are often overlooked are added to this, complicating the issue:

Firstly: thesubstantial difference in part between

a) on the one hand the structure of demands on the armed forces in peace time, which is strongly characterized by elements of bureaucratization, technocracy and pedagogism and which therefore demands administrative as well as psychagogical abilities to a relatively high degree, and

b) on the other hand the structure of demands on the armed forces in wartime, which is constantly determined by the element of uncertainty and therefore to a great extent demands the gift of improvisation coupled with decisiveness and enterprise as well as with general strength of character.

This discrepancy in demands alone makes planning and assembly of a personnel structure "according to functions" in the sphere of the regular officer more difficult.

Secondly: the affinity of modern industrial societies with their armed forces , which influences the quality and quantity of the "rate of new accessions" is subjected, in the Federal Republic of Germany for example, in part to considerable fluctuations.

Thirdly:The individual requirements and understandable professional expectations of the individual regular serviceman are not infrequently controversially opposed to the institutional demands which are often inadequate with regard to society and environment of the armed forces.The build-up of the Federal Armed Forces took place under the political model of the "citizen in uniform", with the basic idea that the profession of a serviceman ought to be "a profession like any other", and therefore ought to be comletely comparable. The contrary, which is indicated here, inevitable leads in part to tension which makes planning according to requirements and the career development according to ability additionally difficult.

The profession of the regular officer is situated within a complex area of tension with numerous negative factors both predominately paramount management aspects as well as general and individual aspects concerned with personnel structure, which lately arise. The "personnel managers" responsible in this field are therefore often confronted with problems the solution of which demands almost the impossible, and indeed: because there is a great deal of uncertainty, because material possibilities and other constraints often permit only too little latitude, because the institutional demands for "the ability to compete", the level of presence and the general ability to function of the armed forces often
a) cannot be fully covered quantatively
b) nor can they be sufficiently "harmonised" with individual requirements as far as quality is concerned.

This often induces short-term expedients or specious solutions. These, in turn, lead mostly to negative after-effects which in the medium and long-term can aggravate the difficulties of the necessary regeneration and motivation. Therefore, taking into consideration the magnitude of the armed forces personnel, ad hoc measures which are of only short-term use are always particularly dangerous because the body of the personnel can prove to be very sensitive in its long-term reactions.

2. On the Situation of the Federal Armed Forces.

The German Federal Armed Forces Association has repeatedly pointed out these connections and others similar to them. Now, for the first time, in the 1979 white paper the Federal Government has openly itemized just such a problem felt by regular officers within the sphere of its negative effects, but indicates no solution.

The extent of the German Federal Armed Forces personnel is 495,000 servicemen in the authorized strength, and of them 270,000 ought to be short-service volunteers and regular soldiers, that is 55%, and 220,000 conscripts. In fact, however, the armed forces have 232,000 conscripts, that is 5.4% too many, in order to compensate for the lack of regular soldiers, which in 1978, for example, ammounted to 8.7%. Ultimately 30,000servicemen are included in this so-called "variable amount",who are not available for the primary work of the armed forces, or are only conditionally available, but who nevertheless but a strain on the budget and personnel structure.

The Federal Armed Forces are equipped with approximately 42,000 officers. Over 31,000 of whom are regular officers, and approximately half of these are field grade officers. The "core" of the armed forces is formed by the regular line officers. Here the Federal Armed Forces have already over 21,000 regular officers at its disposal, although less than 18,000 are needed. 13,000 of the 21,000 regular officers are even field grade officers.

Hence follow the so-called "reassignment backlog and promotion barrier" with all their manifold negative consequences. If, for example, the criteria of a firm of world-wide importance such as Daimler-Benz (Mercedes), which employs 1.3% of its executive in functions comparable with those of a field grade officer, are applied as a gauge, then the Federal Armed Forces, with their present actual strength of personnel of approximately 3%, would, theorectically, be more than twice as well staffed. But such gauges cannot be applied to the armed forces, for example for the reasons given above. Everything should rather be done to at least stem the development of an unbalanced structure which shows a tendency towards propagating itself in a never-ending circle.

3. Consequences of these so-called "reassignment backlogs".

One of the results of the situation presented above is that, for instance, officers born in the years 1935 to 1944 to a certain extent reach higher ranks considerably quicker than planned in the target structure. These officers then remain for a proportionately long time in these service ranks and thereby block the advance of officers from junior age groups to positions commeasurate with their age and qualifications. This results in the so-called "reassignment backlog".

One of the most significant negative consequences of this "backlog"situation is that it leads to a swifly progressing rise in the ratio of advanced age groups. It will be characterised by the following developments if corrective measures are not taken:

Among regular officers there were/will be the following percentages over 40 years of age:
- in 1978 - 42%
- in 1985 - 66%
- in 1990 - 75%.

One particularly drastic retrogression in transfers between all levels of responsibility among regular officers between 1982 and 1991, as a result of the low number of those retiring, will face a no less grave surplus of retirements in the 1990s, namely by the strongly represented age groups. The latter at a time when the increasingly scant age groups characterised by the drop in the birth rate due to the contraceptive pill(from the middle to the end of the 1970s onwards)will be ready for military service, age groups which are so scarcely represented that the number of young men then eighteen years old will not even suffice to cover the conscript target.

As a result of those years with a low birth rate the situation of the labour market, especially with regard to the new generation of executive manpower, will also be under a great strain. To go by experience, this has negative consequences for the number of applications as officer candidates, both with regard to quantity as well as to quality.

The reassignment backlog leads furthermore to a "promotion barrier". This means that, besides the negative consequences it has for efficiency and operational readiness as a result of the high proportion of advanced age groups in the higher ranks such as, for example, company commander and batallion commander , and besides the impeded personnel development for the high and highest ranks, it also affects most officers individually as well. That highly signigicant element of one's working life, job satisfaction, which is based on the assignment of work according to ability, can be attained by making too high demands rather than too low. Stagnation in the working life of the regular officer in the case of the "typical" regular officer leads almost inevitably to a series of "negative" reactions, such as frustration in some cases, or a forced competitive attitude concerned only with their own careers in others. Both extremes are equally injurious to the necessary motivation in the services and to the element of comradeship, indispensible to the armed forces, and which is of decisive importance to functioning ability, especially in times of crisis.

The impairment of career development caused by the "backlog", like that of the career expectations of entire age groups connected with it, inevitably has negative reverberations on the whole career image with the general public.and then leads to new regeneration difficulties.

4. Model of an ideal pyramid as opposed to reality.

By the end of 1977 approximately 1,600 line captains had fulfilled the minimum requirements for promotion to the rank of major. However, only 400 of them are holders of positions as field grade officers.

By 1982 there will foreseeably be a mere 1,000 chances of promotion for captains as a whole. Moreover, approximately 400 candidates for promotion have to be added to this figure annually. These figures roughly characterise the size of the subjectively perceived so-called promotion barrier in the case of captains.

In 1976 the principle for pay according to function was introduced in the Federal Military Pay Act. It sets down the following consequences: every post is to be assessed appropriately; the appropriate operations staff in TO&E (The Table of Organisation and Equipment) is to deal with the assessment in the military sphere. This assessement needs the endorsement of the budget department and that of the Minister of Finance. Authorities in charge of personnel have nothing to do with this assessment. The TO&E is its anticipated demand. The posts plan establishes whether the TO&E assessement is also covered by positions planned in the budget. According to the principle of pay according to function, promotion is conditional upon the transfer to a promotional reassignment.

Only the so-called "changable posts" still offer a limited flexibility.

On the promotion situation with the example of the army : the army has over 4,450 TO&E major posts. They are held by 3,800 majors/lieutenant colonels with 650 colonels. The army has over 3,500 TO&E captain posts. They are occupied by 2,700 captains and 800 majors.

Of the approximately 400 captains in field grade officer positions, only so many can be promoted as majors are transfered from TO&E captain positions to major positions and promoted to lieutenant colonel.

Of the approximately 800 majors in TO&E captain positions only 500 however vacate their planned positions by promotion. In the case of the remaining 300 majors the so-called authorization of 1965/66 is lifted with their transfer. 300 authorized positions (positions as major) will ve dropped by 1982. That means that 300 promotional opportunities for captains will be lost within the next three? years. 420 captains in TO&E major posts who fulfill all requirements for promotion, are therefore already subject to a particularly severe promotion barrier.

The reasons for this are, inter alia, that usually there are too few short-service officers. To balance this out more and more officers have to acquire regular status. In the age groups 1925 to 1934 the number of applicants as regular officers was very inadequate. The consequence of this was the taking over of officers from the age groups of 1935-1944. This resulted in an inorganic age structure among officers, which is additionally burdened with a surplus of approximately 1300 regular officers.

The number of officer candidate appointments is governed by the number of officer retirements. As the manpower levels approved in the budget have been reached and the number of retirements is below normal, fewer officer candidates must be appointed than would be necessary to balance out the structure. This prolongs the eternal circle.

The Budget Structure Act of 1975 has, moreover, increased the actual levels of regular officers by an entire age group.

It is also imperative to find a solution to this problem because of the necessary equality of treatment too. This is made most obvious in this example:

If the work performance of 3 officers is rated with 3 (good) and of them
- one is a captain in a captain's post
- the second is a captain in a field grade officer post
- and the third is a field grade officer in a captain's post,

then this example shows that there is no common denominator for the assessment of work performance. The psychological consequences resulting from this are self-evident.

Now for the so-called ideal pyramid:
It shows 38 age groups with various rank groups. Out of the broad foundation of short-service officers the narrow pyramid of regular officers emerges; the regular officer should be promoted:
- at the earliest at 33
- at the latest at 36 to major
- at the earliest at 36
- at the latest at 45 to lieutenant colonel.

The normal professional goal in this case is lieutenant colonel. The danger of failing the field grade officer examination and of retiring as a captain should be reduced 20 fold.

One out of six lieutenant colonels has the opportunity of becoming a colonel. One out of six colonels has the chance of being promoted to General. The youngest colonel ought to be 39, the youngest brigadier general 45 years old.

The targeted levels of approximately 16,000 line officers will be reached by the actual levels in 1980. However, here there will be approximately 800 regular officers too many, and equally, 800 short-service officers too few. It is now already difficult to control the level of 10,800 regular officers and 4,500 short-service officers in such a way that the posts limit is not exceeded and the increasing number of regular officers does not gradually supersede that of short-service officers. The officer candidate supplementary quota appropriate to the structure of 1,250 has already been choked. The "baby boom" age groups will be taking their school leaving examination in 1982.

In the officer profession age limits are indispensible. For example: company commander/captain 35 (major 40)
- batallion commander 45
- colonel in active service 50.

Irrespective of this, personnel movements within each rank are also indispensible as "lateral transfers". If such transpositions without promotion do not coincide with the aforesaid age limits, then reassignment backlogs occur in the lower ranks with all their negative consequences which then, in turn, alter the structure of the pyramid. Just as detrimental, indeed, even graver, could be the decrease in retirements coupled with a constant number of accessions. Such an extreme situation is imminant at the present time, as a result of the inorganic age structure within the actual strength pyramid:
- those age groups up to 1924 who fought in the war show normal strength
- the age groups without conscription of 1925 to 1934 show a deficiency of approximately 1,500 officers
- and the age groups of 1936 to 1944show a surplus of 2,300 regular officers compared with the authorized strength.

By 1981 normal retirements will be countered by an above normal number of candidates. That means:
- reassignment backlog for captains.
- reassignment drive above and beyond the age groups without conscription and into the older age groups with a surplus representation.

From 1982, however, the number of retirements will remain far below the norm for ten years, but the number of candidates will thereby still be above normal. That leads to
- a "super backlog" in all levels among the ranks.

In 1985 the situation will become extreme, when the age group of 1928 with 64 officers is ready for retirement. Then the positions vacated will be barely sufficient to allow the new accessions to be promoted to colonel and general posts according to the requirements. This situation means:
- for every officer who is promoted sooner than planned in the pyramid, another officer must wait so much longer.

Inevitably a section of the age groups between 1940/1944 will not attain the reassignment as field grade officer before retirement, if no corrective measures are taken. For this group two causes of the backlog situation are particularly applicable:

- the supply of needs for high reassignment levels must remain just as much a priority as the adherance to age limits in certain functions.

Thus personnel management in the 1980s will find itself in an unsolvable dilemma, torn between the efficient staffing of posts and the fulfillement of the justified claims of the individual.

The mounting proportion of advanced age groups in the function which is connected with every reassignement backlog will achieve proportions in the 1980s and 1990s which could endanger operational readiness. By 1985 the proportion of 20 to 40 year old officers will have dropped by 25%. The proportion of 40 to 50 year olds will quadruple. The average age of line officers will increase proportionately. By 1992 every other regular officer will be over 48 years of age. It will become increasingly difficult for this officer corps to maintain the operational readiness of our army's equipment which has been kept up to date by large investments.

Solutions.

As far as solutions to this problem are concerned, only theses can as yet be put forward at this present moment in time. This is itemized in the following six theses:

(1) The backlog problem is above all a problem of age structure, which has developed into a complex series of problems through the absence of pertinent and timely measures, and which has prevailed through to the lower ranks. That is why a sudden pat solution to the entire problem is neither conceivable nor viable.

(2) Short-term, preliminary corrective measures for the 1700 captains at present stagnating in the backlog situation are absolutely imperative with regard to the effects on the armed forces. They must be planned with the goal of signalling movement and of sparking off the first initiative. This can occur as a result of :
- The reduction of special age limits for servicemen by a year
- the budgetary clearing of all appropriate TO&E posts which are not, however, covered by the budget.
- the increasing employment of servicemen in suitable so-called "exchange posts" (that is, posts which theoretically can be staffed by both civil servants and servicemen, but which, in practice, are staffed almost exclusively by civil servants.).

Only a conception effective in the long term for the the lasting solution of the complex of problems in several transitional stages can provide a guarantee for planning and management theoughout the personnel sphere. Such a conception has two main purposes:

- it must make a sensible adjustment between actual and authorized strengths possible,
- it must make it possible to coordinate the imperative staffing of certain functions adequately with the fundamentally equal treatment of all age groups.

(4) The bases for the development of such a conception must be above all:

- an analysis of functions and requirements.
- an appropriate analysis of the satisfaction of needs.
- a mathematical model for the development of a pyramid of actual and authorized strengths to be adjusted in stages on the data basis of age structure-service rank structure - retirement policy - recruiting policy - systemization - distribution of losses - mobility obligations.

(5) After the compilation of bases for the conception, the development of a "supplementary pyramid model for posts" will be imperative, in order to make it possible to suitably bridge the adjustment phases between the actual and the authorized strengths.

(6) The development of such a conception cannot be based solely upon the internal factors of the Federal Armed Forces. Equally it must take into consideration the external factors relevant to personnel structure as well. Otherwise it cannot possibly offer a lasting solution.

The German Federal Armed Forces Association intends to devote itself intensively next year to this entire complex, the solution of which cannot be put off any longer

DEVELOPMENT PLANS FOR AN AIR FORCE OCCUPATIONAL RESEARCH DATA BANK

Robert W. Stephenson

Air Force Human Resources Laboratory (AFHRL)
Brooks AFB Texas 78235

In recent years, the information explosion has affected many different parts of the Air Force. Technical journals and books have increased in number at least tenfold; hundreds of recurring reports are generated by large numbers of staff and headquarters offices; and thousands of xerox machines are busily duplicating millions of pages. Nowhere is this information explosion more overwhelming than in the case of occupational data. Vast quantities of information about Air Force occupations are widely dispersed in many different organizations, with many different formats and degrees of coverage. To add to this, the occupations, organizations, and variables of interest change rapidly with time. Yet there is a growing need for total systems management, total systems research, and longitudinal studies of trends.

The problem is illustrated by the information currently maintained at the Air Force Human Resources Laboratory (AFHRL). AFHRL maintains 29 different kinds of computer files obtained from many different sources; 795 data files on 2034 reels of tape; and more than a thousand technical reports going back to 1943. In addition, AFHRL is the official Air Force repository of all occupational survey data files generated by the U. S. Air Force Occupational Measurement Center (USAFOMC). This consists of 1065 files on 1272 reels, for a total of 391 separate studies.

The problem is not that AFHRL has too much information. The problem is that AFHRL does not have enough information to keep track of all that is going on. Headquarters USAF, the Air Training Command (ATC), the AF Manpower and Personnel Center (AFMPC), the AF Inspector General (AFIG), the AF Inspection and Safety Center (AFISC), and the AF Management Engineering Agency (AFMEA) have their own data bases and generate numerous recurring reports, regulations, studies, and pamphlets. In addition, each occupational specialty has its own functional manager in AFMPC, its own training manager in ATC, its own Specialty Knowledge Test (SKT) monitor in USAFOMC, and its own Pentagon office of prime responsibility. These managers each maintain their own files, and request special reports for their own use. Audiovisual information of many different kinds is maintained by the AF Audiovisual Center. Numerous DoD-wide data bases are maintained by the Defense Manpower Data Center, and distributional data are readily available from remote terminals connected to AFMPC and AFMEA information centers already in existence.

The Air Force Human Resources Laboratory is responsible for conducting research on occupations and manpower utilization in the Air Force. In view of the information explosion that has occurred, AFHRL has several concerns that pertain to its capability to conduct research on occupations. One concern is that HRL has limited access to recurring reports, microfiche, audiovisual information, and certain data bases produced by others. This means that research could be initiated without a comprehensive review of useful information, and that computations previously performed by other organizations could conceivably be duplicated. To make research even more difficult, there are many inconsistencies in the data sets obtained from other organizations, and occupational information at HRL is scattered throughout many tape files.

Plans for an Air Force Occupational Research Data Bank (ORDB)

In order to alleviate the problems created by the information explosion, AFHRL proposes to assemble and consolidate large quantities of recurring reports, digital data, microfiche, and audiovisual information into a centralized data bank, coded for rapid retrieval of information about occupations, with user interaction capabilities, and a large number of tested procedures for conducting different kinds of research studies.

Many different research products will be generated by the ORDB. New occupational ratios, indices, and trend statistics will be developed. Historical files will be restructured in standardized formats that facilitate research. New problem definition methods and predictive models will be designed and evaluated. Relationships will be discovered between sets of data that were never related to each other before. Information needed to design research studies will be readily available in a central location; and data bases needed for the development of new kinds of decision models will be generated.

There are also many practical advantages. When the Occupational Research Data Bank is established at AFHRL, AFHRL researchers will be better informed. Efficiency of computer operations will increase. A few multipurpose computer runs will eliminate the need for large numbers of small scale studies. The Air Force will be provided with a quick response capability for data that are located uniquely at HRL; and duplication of effort can be avoided.

In view of such advantages, the Air Force initiated work on a set of plans for an occupational research data bank in July, 1978. The prime contractor for developing the ORDB is the GSA data processing services contractor for the region, which is currently Kentron International. Consulting advice regarding hardware decisions and overall systems design is being provided by the Naval Ocean Systems Center.

ORDB Development Plans

The ORDB will be developed in three stages over a period of 10 years. Model #1 (to be developed during FY78-82) is a simplified, rapid access enlisted model for high priority items that will include more than 400 summary-description variables for each occupation, as well as occupational survey data for enlisted specialties. Model #2 (to be developed during FY82-84) is an expanded officer-plus-enlisted model that will include model #1 plus over 400 occupational variables describing officer utilization fields. Model #3 (to be developed during FY84-87) is an expanded version of model #2 that will include selected organizational and manpower requirements data, as well as expanded capabilities for cross-tabulations and correlational analyses.

Illustration of Problem Analysis and Solution Evaluation Capability

Figures 1 and 2 illustrate the kind of problem analysis and solution evaluation capability that would be generated by the ORDB. Occupational problems would be diagnosed with the aid of such things as a force

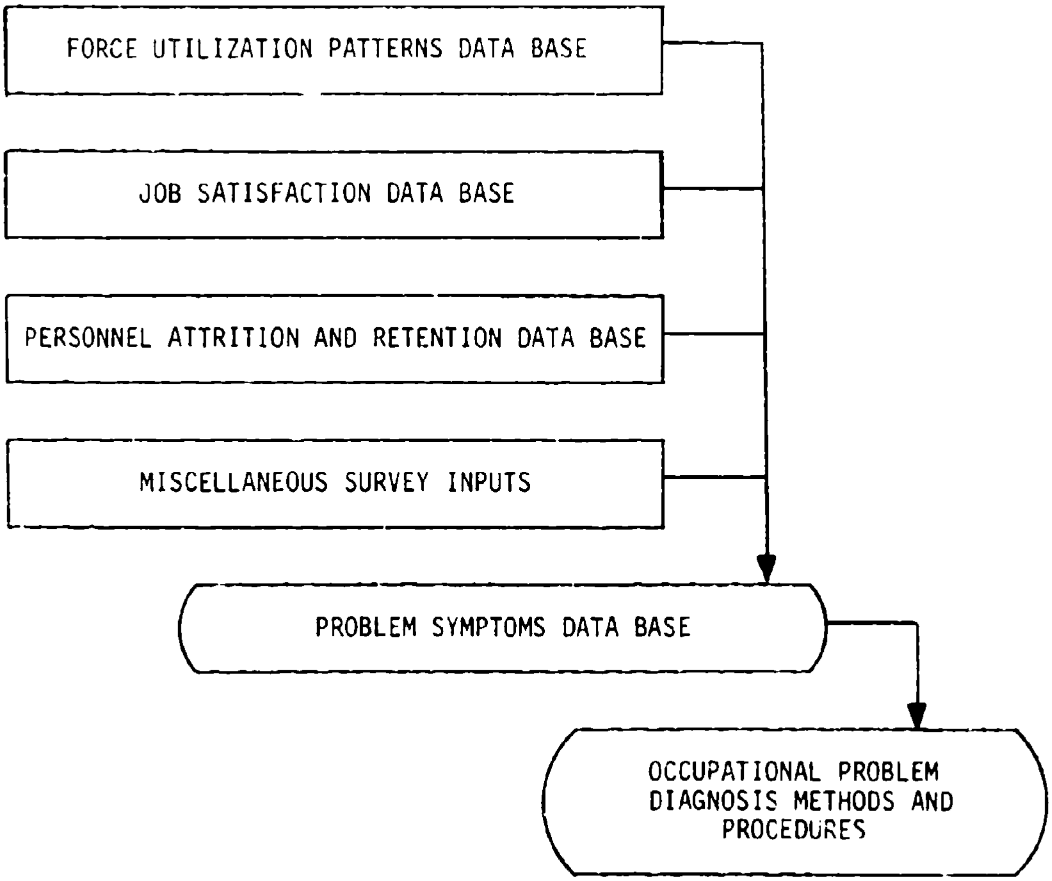

FORCE UTILIZATION PATTERNS DATA BASE
JOB SATISFACTION DATA BASE
PERSONNEL ATTRITION AND RETENTION DATA BASE
MISCELLANEOUS SURVEY INPUTS
PROBLEM SYMPTOMS DATA BASE
OCCUPATIONAL PROBLEM DIAGNOSIS METHODS AND PROCEDURES

Figure 1. One of many possible problem analysis models that could be constructed with ORDB components

utilization patterns data base, a job satisfaction data base, a personnel attrition and retention data base, and a variety of survey inputs. Various solutions to occupational problems would then be evaluated with the aid of other ORDB components (see Figure 2) such as occupational family structuring and job redesign methods and guidelines, force mix decision guidelines, and contingency plans for talent shortages. It should be emphasized that Figures 1 and 2 are only illustrations. Many other problem analysis and solution evaluation data bases and guidelines could have been included in Figures 1 and 2, which is why we need a large multiple-component, flexibly-designed Occupational Research Data Bank.

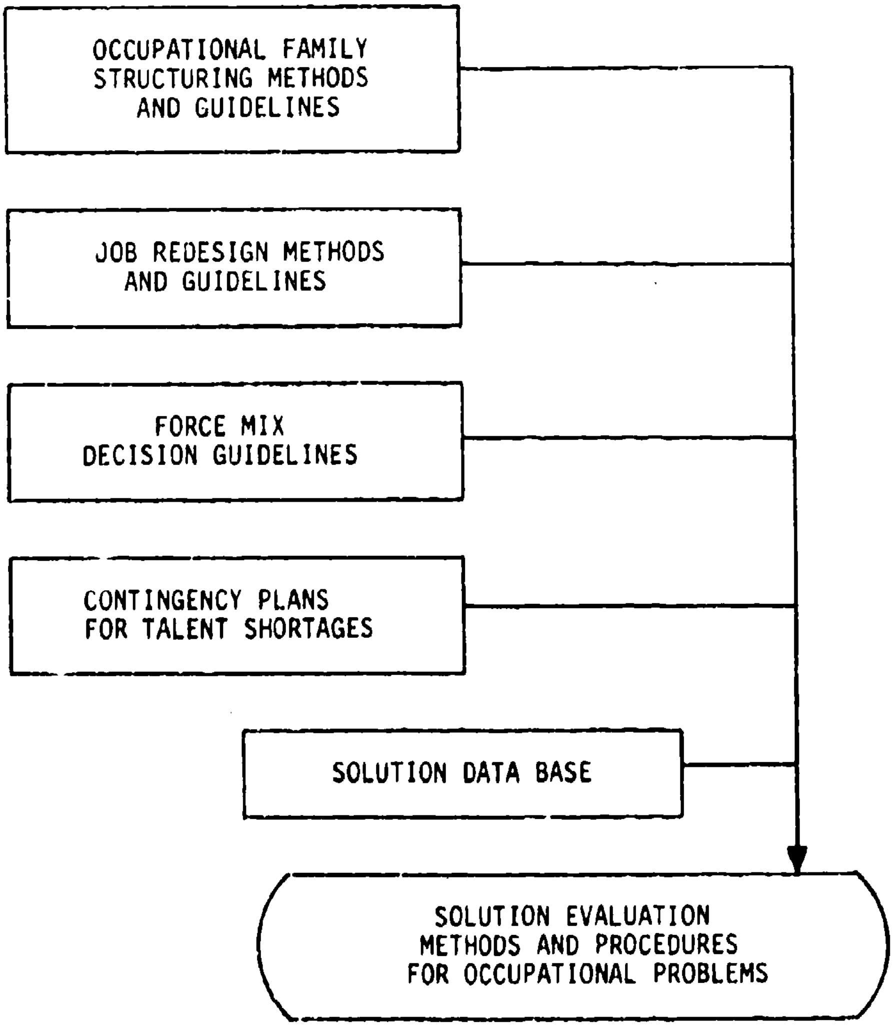

OCCUPATIONAL FAMILY STRUCTURING METHODS AND GUIDELINES
JOB REDESIGN METHODS AND GUIDELINES
FORCE MIX DECISION GUIDELINES
CONTINGENCY PLANS FOR TALENT SHORTAGES
SOLUTION DATA BASE
SOLUTION EVALUATION METHODS AND PROCEDURES FOR OCCUPATIONAL PROBLEMS

Figure 2. One of many possible solution evaluation models that could be constructed with ORDB components

Illustrative Questions and Predictive Models

Tables 1-6 illustrate the kind of questions that could be answered and the kind of predictive models that could be developed. Table 1 contains questions about occupational trends and distributions that would be answered by model #1 (enlisted data only). Table 2 contains examples of predictive research studies that could be conducted using these data sets. Tables 3 and 4 illustrate the kinds of questions that could be asked and the kind of research that could be conducted using model #2 (officer plus enlisted). Tables 5 and 6 illustrate the capabilities of model #3 (officer plus enlisted plus individual plus organizational).

Utilization of ORDB Components to Meet HRL Research Objectives

Priority in the design of the ORDB is being given to certain research products that are needed to support HRL research thrusts. Table 7 is a tentative list of the thrust components to be generated by the ORDB and identifies the AFHRL research thrust in which the ORDB research product will be used. The two job properties and job requirements data bases will be used as components of officer and enlisted assignment systems. The two occupational problem solving research components

Table 1. Illustrative Questions about Trends and Distributions from Model #1 (Enlisted)

Question	Research Application
What is distribution of Combat Skill designations for personnel in each occupation?	Transferability of skills research
What are the trends with respect to percent civilian jobs in each occupation?	Civilian survey plans
Are there any pending or planned equipment or Specialty Training Standard changes in this specialty?	Sample selection decisions and revisions to existing survey instruments
How much time is spent in comparable supervisory tasks by the E-9s in each specialty?	Consolidation of E-9 positions into higher level managerial jobs

Table 2. Illustrative Predictive Models Involving Diverse Sets of Data in Model #1 (Enlisted)

Predictors	To be Predicted
Percent fill in occupation Exposure to temperature extremes Avg aptitude score	Accident rate for each occupation
Exposure to noise extremes Avg overseas tour length Civilian pay opportunities	Attrition or retention in each occupation
Job difficulty Probability of accidents Avg training priority rating for all tasks performed	Course length
Status of occupation Transferability to civilian sector Percent of population that meets aptitude requirements	Recruiting difficulty in each occupation

Table 3. Illustrative Questions about Trends and Distributions from Model #2 (Officer + Enlisted)

Question	Research Application
Which officer jobs are thought to require an engineering degree?	Research on officer education requirements
What are the equivalent civil service grade levels for officer jobs?	Research on civilianization of military jobs
What duties currently being performed by officers could be performed by senior Noncommissioned Officers?	Contingency plans for talent shortages
Which officer jobs seem to benefit most from professional military education of the job incumbents?	Research on officer assignment systems

Table 4. Illustrative Predictive Models Involving Diverse Sets of Data for Model #2 (Officer + Enlisted)

Predictors	To be Predicted
Job Difficulty Weapon system complexity Importance of safety considerations	Course length.
Civilian pay opportunities Officer job properties Officer job satisfaction	Officer retention rate
Average number of duties assigned Average number of extra duty hours Job responsibility index	Officer promotion rate
Probability of casualties in wartime Requirement for military knowledge Requirement for supervision of military personnel	Percent civilians with that Air Force Specialty Code

Table 5. Illustrative Questions about Trends and Distributions from Model #3 (Officer + Enlisted + Individual + Organizational)

Question	Research Application
What is the male/female mix for personnel assigned to different work centers?	Research on utilization of women
Which organizational units have the most accidents?	Safety research
Which specialties are expected to have increasing skill and knowledge requirements during the next five years?	Longitudinal studies of job changes
What is the personnel turbulence rate in different kinds of organizational units?	Research on attrition and retention
Which types of maintenance organizations require the most investment in on-the-job Training?	Cost effectiveness of maintenance organizations

Table 6. Illustrative Predictive Models Involving Diverse Sets of Data in Model #3 (Officer + Enlisted + Individual + Organizational)

Predictors	To be Predicted
Aptitude scores Minority status Professional military education	Career progression patterns in difficult specialties
Exposure to x-rays Exposure to temperature extremes Percent fill of manpower requirements	Probability of hospitalization within different kinds of organizational unit
Minority status High school grade point average Compatibility of assignment with vocational preference	Awards and adverse actions in various commands
Number of authorized flying hours Avg skill level of assigned personnel Type of maintenance organization (POMO vs non-POMO)	Sortie rate during operational readiness tests

will be incorporated into enlisted and officer force maintenance and career progression systems. The task training decision algorithms will be used to help make decisions about "where" and "when" training should be provided. The plans for model #3 are very tentative, but the various manpower and organizational research components should be useful in research on future manpower requirements and organizational effectiveness measurement systems.

Other Utilization Plans

Although the ORDB is primarily justified in terms of research components that support approved HRL research objectives, other uses of a more general nature will exist. One important use is for research design and sample selection. Researchers are sometimes embarrassed to discover that their research is less useful than it might be because they were unaware of changes that had taken place in an occupation or training course. These changes could have been anticipated if the researchers had been better informed. Another important application involves exploratory studies of relationships between sets of data that have never been interrelated before. For example, data on accidents maintained by the AF Inspection and Safety Center have never been related to data on force manning (maintained by the Air Force Management Engineering Agency) largely because the two data bases are widely dispersed and are maintained by different agencies in different formats. Once selected

Table 7. Utilization of ORDB Components to Meet HRL Research Objectives

Model #	Thrust Components to be Generated from ORDB	HRL Research Objective
1	Enlisted job properties and job requirements data base	Enlisted assignment systems
2	Officer job properties and job requirements data base	Officer assignment systems
1	Data for occupational problem analysis and solution evaluation methods	Enlisted force maintenance and career progression systems
2	Data for officer utilization problem analysis and solution evaluation methods	Officer force maintenance and career progression systems
1	Data for task training decision algorithms	Training decision systems
3	Occupational manpower requirements data base	Future manpower requirement forecasting techniques
3	Work center capability prediction models	Organizational effectiveness evaluations

items have been converted into a consistent format and incorporated into the ORDB, it will be possible to ask many interesting questions about relationships between these two sets of variables. One might ask, for example. about the extent to which accidents are related to the percent manning in a work center.

Another use of the ORDB involves historical and longitudinal files. Most organizations have very good data about current problems, but they can rarely answer questions about long range trends because the data are almost never in a consistent format. Policies change, reorganizations occur, occupations are redefined, and the priority is always given to this year's data rather than last year's data.

Another planned use of the ORDB is for orientation of those who are about to conduct research on an occupation. The ORDB will contain vast quantities of information in hard copy files, technical reports, and microfiche. Since all of these will be coded and indexed on a special computer file, the ORDB will help researchers to become familiar quickly with everything there is to know about an occupation.

SUMMARY

Plans are being developed for the establishment of an Air Force Occupational Research Data Bank that will contain hard copy, microfiche, disc, tape, and audiovisual components. The major objective of the ORDB is to support approved AFHRL research thrusts by generating data bases and components that can be used in the design of new occupational problem analysis systems and decision models. The data bank will also be used for trend studies and research design purposes as well as for exploratory research relating widely dispersed sets of data. As a result of establishing the ORDB, the efficiency of HRL operations is expected to increase, and duplication of effort can be avoided. The Air Force will also be provided with a quick response capability for data uniquely located at AFHRL.

ESTABLISHING AN AIR FORCE OCCUPATIONAL RESEARCH DATA BANK

James B. Carpenter, Ph. D.
Wayne B. Archer
Roger L. Camp

Developed under:

AFHRL Task Number ORDB-001
GSA Contract GS-07S-02355
Kentron International, Inc.

INTRODUCTION

An Air Force Occupational Research Data Bank (ORDB) is being developed to provide ready access to a wide variety of current and historical occupational information for research and management use. The prime contractor for this effort is the GSA ADP technical services contractor, Kentron International, Inc. Kentron personnel assigned to this project are co-located with the monitoring activity (AFHRL/OR) at Brooks AFB, Texas.

The original plans for establishment of an occupational information center were implemented in July 1978, when Task OIC (now ORDB) 001 was initiated. At that time, ORDB-001 was essentially conceived of as an information gathering process intended (1) to identify data sources, (2) to develop an overall systems plan for integration of occupational data, (3) to implement a highly automated, multi-media occupational information center designed to serve multiple potential users and (4) to provide multi-remote-terminal access and interactive capability.

The initial effort under this approved task plan centered on surveying potential occupational information sources and center users. The survey was accomplished during the latter part of 1978 and included; various directorates of the USAF Manpower and Personnel Center, Air Training Command, Recruiting Service, Management Engineering Agency, Occupational Measurement Center, Data Services Center and the Air Force Academy. Additionally, the Army Research Institute and MILPERCEN, the Naval Personnel Research and Development Center and Bureau of Naval Personnel, and other DOD and civilian activities, either maintaining or functionally employing occupational data were included in the survey. Conclusions may be summarized as:

[1] The opinions or assertions contained herein are those of the writers and are not to be construed as official or reflecting the views of the United States Air Force.

1. Numerous occupational data bases and related reports are available. Much data is currently available in a quantifiable format within AFHRL data files, but not easily or immediately accessible.

2. A major investment in the equipment required to establish an independent information center is unneccessary at this time. Existing computer facilities (with dedicated storage capacity) can fulfill initial requirements.

3. The role of the center should be redirected primarily towards providing AFHRL with immediate access to a wide variety of current and historical occupational data in a format specifically designed to facilitate research and managerial use. Immediate attention should be directed towards establishment of a quantified research data base with a de-emphasis/deferral of the information/educational applications contained in the original occupational information package concept.

4. The research "pay-offs" from an occupational research data bank are even greater than originally thought. In fact, the urgency of establishing a preliminary data base which will be operational even though limited in scope, is sufficient to justify redirection of the project towards this goal.

Consequently, revisions to the task plan were approved in January 1979, and the Government Project Manager's (GPM's) official long range plans for the total effort were redefined during the following March. These changes modified the original concept and resulted in immediate emphasis on generating expansion over time as data of value is identified/obtained, and postponing complex data development (including inherent equipment acquisition, i.e., computerized microfiche, mini-terminals, etc.) to a later time frame. Task ORDB-001 was redefined as "Information Gathering Survey and Design of File Structure" for Model 1, "a simplified rapid access model for high priority items to include a large number of occupational data variables describing enlisted specialties, as well as occupational survey data for enlisted specialties."

DESIGN FEATURES

The Occupational Research Data Bank (ORDB) file will be structured to meet the following objectives:

(1) Provide the capability to rapidly retrieve information via CRT,

(2) Efficiently utilize storage space,

(3) Maximize the capability to expand the data base,

(4) Take into consideration the nature of the ORDB data, which is primarily historical, large volume, narrative and statistical information on Air Force specialties, and

(5) Simplify maintenance of ORDB software as much as possible.

Chart 1 shows the proposed organizational structure of the ORDB and the developmental procedures through which the objectives will be achieved.

ORDB Specialty descriptive information, both narrative and statistical, will reflect the classification structure of the Air Force and the population subgroupings that are of key interest to the AFHRL. Data will be stored and retrieved at four levels of aggregation for five subpopulation groupings. The levels include Career Field (for example, 43, aircraft maintenance), Career Ladder (for example, 431X0, helicopter maintenace, AFSC (for example, 43180, helicopter technician) and Air Force total airman (across all specialties). The population subgroups included 1st-Term, 2nd-Term, Career, 1978-Input, and Totals, within specialty and across all specialties for the Air Force as a whole. Chart 2 shows these key structural groupings for which the ORDB will provide information.

The ORDB file structure will relate to, but not duplicate, the structural groupings represented by Chart 2. The ORDB file will contain records by specialty at three different levels, and for five population subgroupings. The three levels are Career Field, Ladder and AFSC. Subpopulations at Career Field and Ladder include 1st-Term, 2nd-Term, Career, 1978-Input, and Total. The five subgroups at AFSC level are the combined 1 and 3 skills, 5 skill, 7 skill, 9 skill, and Chief Enlisted Manager (CEM) skill levels. Use of a "Dummy AFSC", such as 00000, will provide Air Force Totals at each of the three levels for each subpopulation, without the necessity of defining and storing an additional type of record for overall Air Force statistics. In addition to making Air Force-wide statistics available for each of the subpopulations, including skill levels, this approach will simplify design of the system.

The ORDB Master File will be stored on disk with back-up copies on magnetic tape. Other data will exist in hard copy, microfiche and audio visual form. An index to these data elements will be included on disk.

OCCUPATIONAL DATA BANK CONTENT

The content of the records is currently in a state of change as new information sources are investigated. However, each record will include three basic types of variables, including (1) Narrative blocks of data for non-quantifiable variables, such as duty description, mandatory training courses, reference, or other information, (2) Indicator type variables such as physical "X" factor requirements, or a flag indicating whether or not a specialty is authorized for award to enlisted women, and (3) Statistical Counters reflecting frequency distributions, means and standard deviations. Most data will be captured, computed, and stored in Ladder level records with fewer variables at the Career Field and AFSC skill level records than at the Ladder level.

The records and their content are Air Force specialty-oriented with emphasis on the Career ladder level records. This approach reflects the AFHRL research emphasis placed upon analysis of Air Force Career ladder characteristics. Likewise, the population breakouts are those specified by the GPM and other ORDB-users as the subgroups of most concern to AFHRL research as well as to high-level Air Force management. In addition, separate records for population groups will allow retrieval of any variable by population, without having to preidentify variables for population breakout at time of data capture. Specialty-oriented records will not duplicate existing AFHRL files containing data on individuals. Appendix 1 provides a listing of representative occupational data variables which will be included in the Occupational Research Data Bank.

SUPPORT PROGRAMS

The approach to date has been toward a table-driven system to permit the easiest possible expansion of the data base resulting from the identification and addition of new variables. In essence, a table-driven system contains one or more tables which describe the record layouts and variable formats of the file. Update and retrieval processes use the tables to determine variable starting record positions, number of characters, etc., to locate, update, and retrieve data in the file. The advantage of this method lies in the ability to redefine or expand a file by updating a table rather than modifying and recompiling the programs themselves. This approach will also simplify handling files for different years, such as airman specialty files for 1978 and 1979, which may include different variables resulting from a change in personnel data maintained from one year to the next. Each file can be described by different tables rather than different programs.

Recent emphasis has been placed on using System 2000, a data base management system, for systems maintained at AFHRL. There are three advantages to System 2000 in relation to the ORDB project. First, using System 2000 in developing the data base and retrieval systems will simplify transfer of ORDB maintenance functions from the contractor to AFHRL in-house resources upon completion of the contract. System 2000 will be common to other systems at AFHRL. The operation and maintenance of the ORDB could be more readily assumed by AFHRL personnel than if it were totally unique. Second, development time may be reduced since less effort will be needed to design the data base. The third advantage stems from System 2000's interface capabilities with FORTRAN and other languages. As the ORDB evolves, it will be more easily restructured under System 2000 to interface with statistical routines available in FORTRAN for example, or to take advantage of other user modules already developed in-house.

SUMMARY

At present, contract work efforts are being concentrated on the collection and generation of occupational information at the specialty level from readily accessible data files. Primary emphasis is being directed towards AFHRL's extensive historical files of Air Force enlisted personnel records and the results from occupational surveys of airman career ladders (CODAP data). Systems design work is continuing together with identification of additional information sources. The first operational model of the ORDB, with restricted capabilities, is scheduled for completion in June 1980. This model will provide user access to several hundred variables on all airman specialties with relatively simple retrieval and display capabilities. However, variable selection for inclusion in the first prototype has been governed by both accessibility and historical and projected AFHRL research thrusts in occupationally related areas. These concerns suggest that the data base should be structured to support both descriptive analysis for identification of existing cross-specialty differences and longitudinal analysis for identification of changes within occupational specialties.

Advanced features including expansion of the data bank and improved output and analytical capability will be added in later work steps resulting in a data bank of approximately 400 summary-descriptive variables and integrated occupational survey data for each enlisted specialty. Model 1, as described, is scheduled for completion during FY 82.

Model 2 of the ORDB will add occupational variables describing officer utilization fields and is projected for operational implementation during FY 84. Model 3 will incorporate additional occupationally-related data including selected individual, organization, and manpower requirements data together with expanded display and analytical capabilities as identified during the development of earlier models. This model, envisioned as an on-going permanent facility, will essentially complete the required system and is projected for implementation in FY 87.

Chart 1
ORDB System Overview

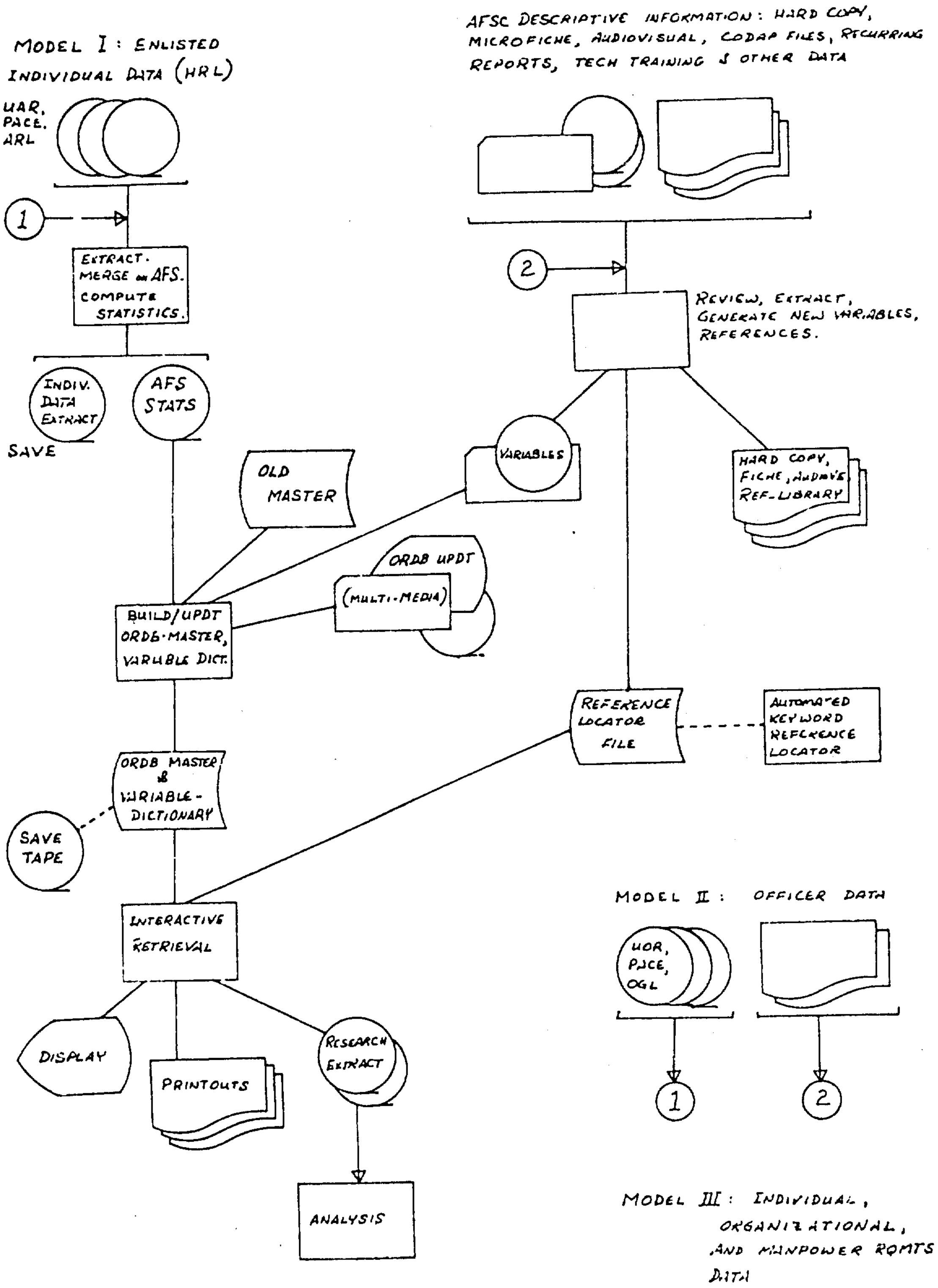
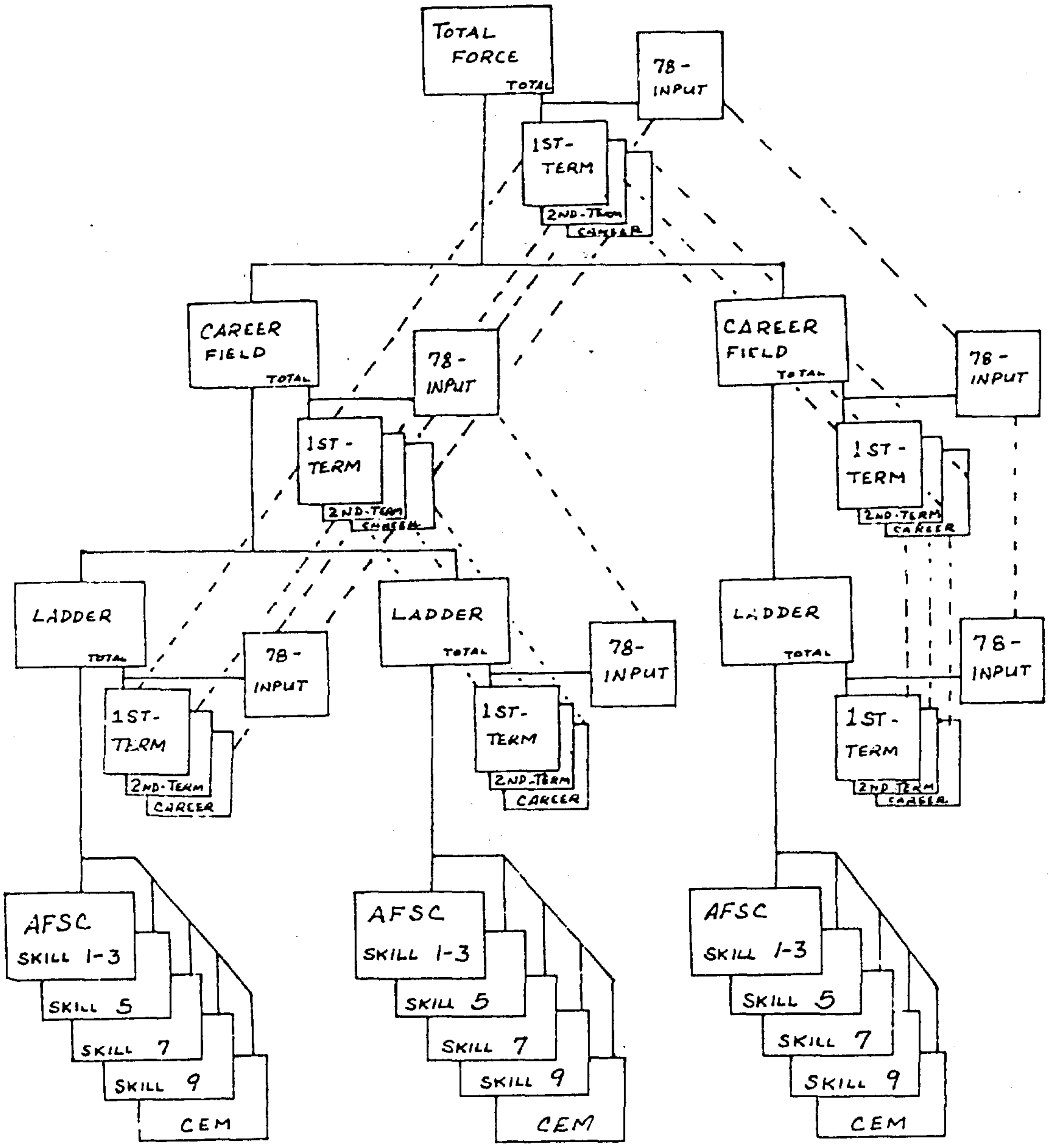

MODEL I: ENLISTED INDIVIDUAL DATA (HRL)
UAR, PACE, ARL
1
EXTRACT-MERGE ON AFS. COMPUTE STATISTICS.
INDIV. DATA EXTRACT
SAVE
AFS STATS
AFSC DESCRIPTIVE INFORMATION: HARD COPY, MICROFICHE, AUDIOVISUAL, CODAP FILES, RECURRING REPORTS, TECH TRAINING & OTHER DATA
2
REVIEW, EXTRACT, GENERATE NEW VARIABLES, REFERENCES.
VARIABLES
HARD COPY, FICHE, AUDVS REF-LIBRARY
OLD MASTER
ORDB UPDT
(MULTI-MEDIA)
BUILD/UPDT ORDB-MASTER, VARIABLE DICT.
REFERENCE LOCATOR FILE
AUTOMATED KEYWORD REFERENCE LOCATOR
ORDB MASTER & VARIABLE-DICTIONARY
SAVE TAPE
INTERACTIVE RETRIEVAL
DISPLAY
PRINTOUTS
RESEARCH EXTRACT
ANALYSIS
MODEL II: OFFICER DATA
UOR, PACE, OGL
1
2
MODEL III: INDIVIDUAL, ORGANIZATIONAL, AND MANPOWER RQMTS DATA

Chart 2
ORDB Information Structure



APPENDIX I

OCCUPATIONALLY RELATED VARIABLES

1. <u>DEMOGRAPHIC</u>

Age
Education
Minority group
Enlistment region

2. <u>SPECIALTY RELATED</u>

<u>Prerequisites</u>

Aptitude level
Training requirements
Physical profile
Security clearances

<u>Job Descriptions</u>

AFR 39-1 Specialty descriptions
CODAP task-level description
Job-type clusters
Job difficulty index

<u>Attitudinal/Job Satisfaction</u>

Re-enlistment intent
Expressed job interest (CODAP)
Attrition rates
Occupational Attitudes Inventory

<u>Job Properties</u>

Environmental properties
Strength & Stamina requirements
Job hazards
Health/Medical data

3. <u>ORGANIZATIONAL</u>

Base locations
Geographical Areas
Organization charts
Work center descriptions

4. <u>INDIVIDUAL</u>

SSAN/Identification
Aptitude test scores
Training courses
Time in grade

5. <u>MANPOWER</u>

Percent fill
Recruiting difficulty
First term/career ratios
Tour length/rotation indices

CODAP: A CURRENT OVERVIEW

Michael C. Thew
Johnny J. Weissmuller

Air Force Human Resources Laboratory
San Antonio, Texas 78235

INTRODUCTION

Within the United States Air Force, as well as many other military and civilian agencies, there is a fundamental technology which supports both the operational and research occupational analysis programs. The core of this technology is called CODAP, an acronym for the Comprehensive Occupational Data Analysis Programs. The CODAP "system" is a set of analysis tools and procedures which use, as raw material, information provided by the members of the occupational field being studied. This system may be used to revise classification structures, assess job related skills, verify the relevance of training courses and a host of other applications in which an accurate knowledge of job content at the task level is desirable.

This paper presents an overview of the CODAP system designed to familiarize a new occupational analyst with various perspectives on the system and provide a broad-brush view of a typical analysis being translated into the necessary sequence of computer programs. For each of the three perspectives discussed, a basic definition is provided, followed by an analysis highlighting the important concepts and ideas. These perspectives were chosen because they represent the three stages of development in the history of the CODAP system as detailed below, and each is identifiable with a current, on-going effort.

In the mid-to-late 1950's the United States Air Force began to commit resources aimed at the goal of developing and implementing a technology to accurately, reliably and economically define Air Force jobs as a starting point for several personnel management objectives. This initial thrust began with a survey of the then available techniques and rapidly approached, met and advanced the state-of-the-art. Many possible alternatives were tried. Some approaches were suggested based on one theory, others on another, and still others on a third. The one firm conclusion reached, however, was that any proposed technique had to be empirically tested and evaluated, no matter how strongly supported in theory. As basic questions began to be answered, analysis procedures and guidelines began to develop to support and implement the research findings. Not long thereafter, the introduction of the computer extended the horizons on this endeavor, and what had been limited to a research project eventually spun-off a very promising operational concept. The research approach, the operational procedures and the computer programs all represent different aspects of the current system, and each defines a different vantage point from which to view CODAP.

The Research Approach

CODAP: 1. An approach in occupational analysis which stresses the quantification and empirical testing of occupational factors over a defined inventory of items. These items, usually task statements, along with background questions are used to define categories, form groups or clusters and produce prioritized lists meaningful to managers.

The term "CODAP" was originally developed as an acronym to describe a given set of computer programs, but in common usage, it has grown to take on additional meanings. In its purest form, the CODAP research approach is nothing more than the scientific method. At this level of abstraction, the term "CODAP" means an integrated and evolving system of principles, procedures and programs. This viewpoint is necessary to explain why many new procedures and programs are developed in support of CODAP analyses and yet only a few are ultimately accepted as part of the system. At the base of this viewpoint is the desire to generate the most highly useful information that may be obtained at a reasonable cost.

Because of this pragmatic outlook, the purpose of the research approach is to define a conceptual system by which the scientific method may be applied to the field of occupational analysis. Heavy cultural exposure to the basic terms and ideas used to describe the workplace make it hard to consider such ingrained ideas as mere theory, but, in order to produce a computer-based model, it is necessary to redefine common terms in a more technical sense and precisely identify their interrelationships. To use this computer model with the scientific method, the objects defined in the system must correspond to something in the real world such that when predictions are made, they can be tested for their accuracy outside of the computer-model context. In most cases, accuracy is assessed by having managers review the products and form a majority opinion. In the final analysis, customer satisfaction, tempered by a cost analysis, is the basic evaluation criterion.

The computer model is formulated on the premise that the data collected and processed should be quantifiable, verifiable, summarizable, comparable and restructurable. Each of these is a necessary ingredient to insure the utility of the system produced. If the data are not quantifiable, prioritized lists cannot be made because "objects" cannot be compared in an automated procedure. If the data are not verifiable, the system lacks credibility and reliance on such data, by management, may have serious ramifications. As it is important to present managers with useful information and not just mountains of data, it is essential that the data are summarizable in a manner that makes sense to the managers. For this reason, the data must be restructurable to meet the needs of different functional managers and it must still be comparable, even under the new structure, to permit prioritized lists.

Given the overall needs of the Air Force personnel system, the analysis of jobs was chosen as the primary focus for CODAP. In this model, the technical definition for a "job" is a set of tasks selected from a standard inventory. The selection of tasks is done by the job incumbent to reflect those tasks actually performed by the incumbent. In addition to identifying the tasks, the job incumbents are asked to categorize the relative amount of time spent performing each task. These "time spent" categories are typically presented as nine levels with the midpoint being defined as "average." When these "time spent" categories are assigned numerical weights and converted to percentages, individual job descriptions are produced. Composite job descriptions, formed by averaging corresponding percentages for a selected group of incumbents, correlate acceptably well with the more costly "stopwatch" time estimates, but are not directly convertible into workhours (McFarland, 1974). In a series of studies, dating back to the earliest research, direct estimates of absolute time spent have led to such impossible conditions as some employees working more hours per week than were available in a total month. It is apparent that some adjustments must be made before this technique can be used in an operational mode, and research in this area is continuing. With the current technology, however, the increased precision obtained by using more than nine categories does add some to the accuracy, but it is obtained at a considerably higher cost in the data collection and automation phase. In addition, experience has shown that many requirements for precise workhours may be addressed in sufficient detail by using these time-ranked approximations, simply by reworking the statement of the problem to be solved. Solving real world problems with this computer-based model, and extending that model when necessary, is the heart of the CODAP research approach.

This basic computer model has been expanded several times and will probably continue to be enhanced as new research questions arise. Where the original version of CODAP simply reported the time-ranked job descriptions for a specified group of individuals, the second version introduced hierarchical clustering which aided in identifying those types of jobs which actually existed as opposed to those traditionally thought to exist. This technique provides valuable feedback to the management of large or geographically separated organizations. With the jobs thus identified, outdated classification structures and potential duplication of effort may be spotted and corrected.

The next major advance in the system was to incorporate a method of integrating the priorities of supervisors. Just as job incumbents are asked to provide time-ratings, supervisors are asked to provide priority levels. These priority levels may deal with such occupational factors as training time to reach proficiency, consequences of inadequate performance or the criticality of immediate performance. By collecting these ratings from supervisors, processing them in the interrater reliability program to remove non-cooperative raters and computing a composite vector, a prioritized list may be produced which represents the majority opinion of supervisors surveyed. Complex factors such as "importance" or plain "criticality" are addressed by identifying those single-factor components which are thought to contribute to that high-level concept and developing explanatory equations through policy-capturing techniques. This methodology, employing interrater reliability and regression analysis techniques, allows the traditionally people-job oriented CODAP system to deal with those topics currently called Task Analysis; this area is addressed within the CODAP system by the Task Factor programs. In historical documents, "Task Analysis" was used to describe what is now called "Job Analysis" and the modern reader needs to be aware that Instructional Systems Development (ISD) research changed the meaning of this term.

The most recent advance in the system is the ability to recategorize and summarize the task-level information into higher-level modules more meaningful to managers. The most successful application to date has been in identifying those tasks associated with each mandatory training module and reporting, by module, the total time actually spent in that area by job incumbents in the field. Curriculum design and validation will be greatly affected by this application. This technique is not limited to the functional area of training. Any functional area which can define their topics of interest and provide a working-group of subject matter specialists to categorize task statements under those topics can make use of this new technology. Just as the introduction of supervisory priorities increased the utility of the occupational data for supervisors, this advance is the first step in increasing the value of occupational data for the managerial level.

The next major advance to the system will be called Profile Analysis. This enhancement will permit the hierarchical clustering of people or jobs based upon any data items of interest. Potential applications include studies of job satisfaction profiles across jobs, clustering jobs to identify job-related requirements to address assignment questions, clustering supervisors to determine if different priority policies are at work, and perhaps, even clustering of tasks to develop empirically determined training modules/duty-assignment modules. If not part of this enhancement, at least concurrent development will take place to provide for managerial modules, the same review and reporting capabilities as exist for supervisory priority ratings. When these managerial priorities are incorporated, they too, may be clustered and analyzed to identify all the policies governing the allocation of resources within the functional areas covered. As can be seen, this is a dynamic system and the research-approach viewpoint is necessary to understand the evolving nature of CODAP.

The Procedures

CODAP: 2. A set of operational procedures and guidelines used in conjunction with the CODAP programs. These procedures cover areas such as survey instrument construction, survey administration, data processing support, and occupational analysis techniques covering both computer products and external sources.

As research validates new techniques, procedures and guidelines are developed to support those advances in the operational mode. Procedures are devised not only for each of the areas listed above, but also for the interactions between those independent systems. The most comprehensive treatment of these interactions is available in the "Procedural Guide for

Conducting Occupational Surveys in the United States Air Force" (Morsh, 1967). The impact of the optical scanning technology on the data processing and analysis systems was discussed in "Data Processing Problems in Collecting Job Survey Information" (Weissmuller, 1976). Three documents are currently being developed under contract. They include a task inventory construction handbook, an executive overview identifying potential applications and benefits, and finally, an occupational analyst manual. The first will update and expand the ideas in the "Procedural Guide." The second will probably replace the Navy-symposium paper presented by Dr Christal (Christal, 1974) as the applications oriented overview. The third, the analyst manual, will address a long-standing need for a look at the interaction between the requirements of an occupational analysis and the computer processing necessary to support those requirements. The remainder of this section provides a brief overview of that interaction to serve as a starting point for that manual as well as for new occupational analysts.

The CODAP package contains over 50 computer programs which serve the occupational analyst by providing the information necessary to make decisions about an occupational career field. In the operational mode, heavy emphasis is placed on identifying the distinct types of jobs found in the career area and on reporting the majority viewpoint identified in the analysis of the supervisory priority ratings. This parallel interest in incumbents/jobs and tasks/occupational factors is reflected by a symmetry within the CODAP system and will be used in the following discussion. Because the interface between data processing and occupational analysis is being discussed, the perspective from which this topic is approached is that of a neutral or third vantage point. The five categories of "processes" presented are intended to serve as generic labels in the sense that any given computer program or occupational analysis objective will probably require one or more categories to fully explain the program or the objective. In describing the details of these processes, cryptic names will appear within parentheses. These names, specified in all capital letters, represent the computer identification of the CODAP program being described. For further information on that program, refer to Appendix A in which all the programs have been listed in alphabetical order based on those computer IDs.

There are five basic processes which will serve as a basis for this data processing/occupational analysis overview of CODAP. These categories are:

1. Data Preparation and Validation
2. Selection of Individuals to Form Potentially Useful Groups
3. Computation of Summary Information
4. Comparisons of Summary Information, and
5. Prediction

As each process is discussed, the reader should be aware that, in general, each process may have a set of Task Analysis programs and a set of Job Analysis programs as well as a program or two which interact these two main streams.

1. Data Preparation and Validation

Once the job inventory has been administered, the first step required, regardless of the analysis to follow, is to prepare the data. This usually entails automating the data in one or more ways (optical scan and/or keypunch), insuring that the multiple methods produce CODAP processable records (local utility programs for sorting, matching and merging, SETCHK), generating the appropriate computer files (INPSTD, FACSTD), and finally, validating those computer files and producing necessary reference materials (JOBSXX, TASKXX, DICTXX, VARSXX). The most critical audit step is a review of the job descriptions computed for all job incumbents on the computer file (JOBSXX). A check for reasonableness and possible formatting errors should be made. A typical reasonability check is performed by locating those tasks known to be common throughout the occupational area being surveyed and insuring that a high percent members performing value is associated with those tasks. A cursory review of the less-common specialist tasks should reflect a correspondingly low performance level. Format errors may be suspected if one or more of the first or last task items listed in inventory order are found to have no members performing. If any of these tests fail, a potential problem is indicated and a more detailed audit is warranted. Less critical, though more time-consuming checks include reviewing task statements (TASKXX) and background item descriptions (DICTXX) for spelling errors. The documents produced to check spelling also serve as reference material in the request of additional products. As missing responses and incorrect input can be very costly to correct later in the analysis, distributions of background information should be studied to identify inconsistencies and potential problem areas (VARSXX). The document produced to perform these checks is extremely valuable in preparing requests to identify special groups (JOBSPC) of incumbents or in setting limits for the variable summary programs

2. Selection of Individuals to Form Potentially Useful Groups

Because the computer-model on which this system is based accepts only individuals as input, all groups to be studied must be developed and identified by selecting the members to be included in each group. In task analysis, this has simply meant identifying those supervisors who contributed to the majority opinion (REXALL). In some career fields, a clear consensus does not appear for certain occupational factors. For those instances, a new, more powerful inter-rater reliability program is being developed. In this program supervisors may be categorized and grouped based upon any logical combination of their background response information (REXSPC). This method of membership specification is called the special sample selection process. This function has an analog in job analysis (JOBSPC) by which groups may also be defined based on the background items, examples being grade level, time in service or geographic location.

One goal of a CODAP analysis is to empirically derive a classification structure without reference to the current structure which is identifiable from background questions. In this situation it is desirable to divide the input population into a small number of mutually exclusive groups in which the members in eaoh group are as homogeneous as possible with respect to some measure of commonality in their task responses. This is accomplished in the CODAP system by first comparing the task responses of each case to those of every other case (OVRLAP) and computing some measure of their similarity. Next, the two most similar individuals are identified and combined to form a new composite group which replaces the two individuals. This process is continued in an iterative manner, combining two individuals and/or previously defined groups at each stage. This process continues until only a single group remains. This technique is known as the hierarchical clustering procedure (GROUP, KPATH). At this point, the results of the clustering are studied to select those stages at which potentially distinct types of jobs were identified (PRTVAR, DUVARS, GRMBRS, DIAGRM). Once these groups have been identified for further study, it is necessary to compute additional summary information on which to base final judgments. Occupational analysts refer to this procedure of identifying distinct types of jobs as "job-typing."

3. Computation of Summary Information

Summary information, which will be the basis for prioritizing lists and comparing groups may be computed for each task, module, individual or group. When a summary value is computed for each task in the inventory, the resulting vector is called a task factor. Task factors may be computed from either rater data or incumbent background data as a function of those tasks performed. In the same program which identifies a consensus group of raters (REXSPC), three task factors may be produced. These factors are the mean rating, the standard deviation associated with that mean, and the number of raters actually rating the task. The task factors that are computed from incumbent data usually represent the average value for some background variable for all group members who perform that given task. Two versions of that averaging algorithm exist. The first is the standard average value (AVALUE) which is typically used to compute the average time in service of members performing. The second program attempts to correct for very large differences in the number of people who appear at each level being averaged (AVGPCT). This is typically used to compute the adjusted average grade level for each task. These task factors, along with others, may be combined, either mathematically or logically to produce composite factors (COMGEN,FACGEN,FACPRE). The distribution of values in any of these task factors may be reported in histogram format (PLOTIT) along with mean and standard deviation data.

A composite job description, one of the most widely used summaries, are computed from incumbent time spent data for the groups identified either in the special sample selection process (JOBSPC,TSKNDX,AVALUE,AVGPCT) or in the analysis of the clustering results (JOBGRP,JOBSPC). Although the composite job description was developed long before the task factor extension to CODAP, the integrated manner in which the system was expanded permits each job description to be used as three independent task factors. These factors, which also are reported whenever a job description is printed (JOBPRT,JOBGRP,JOBSPC), include the relative percent time spent by all members, the relative time spent by only those members performing, and the percent of members performing. When viewed as task factors, these vectors may be used in the same manner, and in conjunction with, the other types of task factors described above.

The computation of summary information for task modules is relatively new but it does allow any task factor to be summarized in a number of different ways (FACPRT). For example, time spent is normally reported by summating all the values for the tasks in that module while task difficulty is summarized by computing the mean value of all nonzero entries and recommended training emphasis is summarized by computing the mean value including zero entries. The attribute of the task factor, combined with the purpose of the analysis, will dictate which summarization should be used. This method of summarization is predicated on the fact that the module definitions have been established, mapped into the occupational inventory, automated and audited (MODSXX).

Although distinguishing between groups may be the primary goal, it is sometimes necessary to summarize information for each individual such that the newly created variable may be further summarized to develop group-level values. When computing summary information for individuals, it is common practice to interact each individual's time ratings with a given task factor (VARGEN,PROGEN). Examples of this include computing the average learning difficulty of tasks performed weighted by the time spent on those tasks, measuring an individual's similarity to a selected job description, and computing the percent of an individual's time spent on an identified category of tasks. Individual summary informatio may also be developed independent of task factors. Categorical variables may be generated based upon a decision table type logic (PROGEN) or by reference to previously identified groups (MEMVAR). Continuous variables may also be formed by mathematically interacting a series of other variables as is typically done in a regression equation (PROGEN).

The computation of summary information for groups, other than task factors, usually is performed by reporting a frequency distribution table for those items of interest in the incumbent's data record (VARPCT,VARSUM,DIST2X). In those instances where the item is numeric, means and standard deviations may be reported and saved for future comparisons between groups as described in the following process.

4. Comparison of Summary Information

Comparisons of groups and/or factors are accomplished in typically one of two ways. The first method is that of comparing groups/factors by detailing the differences at the task level. The most common example of this method is simply reporting several factors side-by-side, each representing a group of incumbents or raters (GRPSUM,FACPRT). By restricting the comparison to two groups or factors, task descriptions along with the appropriate data may be reported in a manner which highlights those tasks demonstrating the greatest differences (GRPDIF,FACPRT). This same comparison may also be displayed graphically (CURVES) in a scattergram plot.

The second method of comparison relies on the computation of summary statistics. The summary statistics may be further categorized as those which are direct summaries of differences across all tasks and those which are summaries across all members in a group. Those programs which summarize individual differences across tasks present their overall difference/similarity measures in either a tabular, matrix format (MTXPRT,FACCOR), or restrict their attention to selected pairs of groups (AUTOJT,CURVES) and report out one or more measures for each pair. There are two main categories for those programs which summarize across cases within groups; those which provide side-by-side reports (VARPCT,VARSUM) and the one which orders the groups based on the value of those statistics (PLTVAL) and presents the information graphically.

5. Predictions

Although the actual process of making predictions and evaluating them is nearly always accomplished in the context of a research project, the results of this process directly affect the operational procedures. The correlation and regression techniques employed are designed to insure that the most cost-effective method is used in integrating supervisory priorities into the operational program. This is accomplished by demonstrating that either the information necessary to address a new priority factor is already available or that a single new factor may replace several established factors currently being collected.

In general, correlations and regressions are used within CODAP to provide explanatory equations capturing the policy of a large number of supervisors. These policies may cover the prioritization of incumbents/jobs (CORREG) or tasks/occupational factors (FACCOR,CURVES). In attempting to explain a complex factor, such as "importance,"it is sometimes cost-effective to predict this factor from other factors currently available, although they are not expected to fully cover all aspects. By using the information about where this partial model is least accurate, one may guide the search for the missing components by generating predicted values (PROGEN,FACPRE). Reports may then be produced (PRTVAR,FACPRT) which highlight the areas requiring further study. By evaluating the characteristics of the cases or tasks in those areas, one might form a hypothesis as to what factors need to be addressed.

The Programs

CODAP: 3. Acronym for the "Comprehensive Occupational Data Analysis Programs." A set of computer programs used to automate, process, organize and report occupational data.

The CODAP system was originally developed by the Air Force Human Resources Laboratory and has been continuously updated and enhanced by AFHRL. Differing versions of the system are available for use on UNIVAC, CDC, and IBM compatible computer equipment. The United States Air Force uses the UNIVAC version of the system. Since its inception in 1960, CODAP has been continuously expanded by AFHRL, and today's system contains more than fifty computer programs.

The development and refinement of computers have greatly enhanced the size and scope of research projects in the field of occupational analysis. The current package has the capability of processing 20,000 cases, 1700 task ratings per case, and 6000 characters of background information per case. Processing requirements for cases, tasks, and background information warrant a brief discussion.

As noted above, 20,000 cases may be processed with CODAP. The only exception to this rule is that the clustering process is limited to sample sizes of up to 7,000 cases. For analysis purposes, a method has been developed to select subsamples. The first step in creating subsamples is to determine the membership from which to select the cases (JOBSPC/JOBGRP). Then a subsample may be created from either all the members identified (SUBSET) or a stratified random sample of the cases (RANSEL/SUBSET). An example of why this might be done is that most surveys are administered to a very large number of job incumbents within a career field and many times a second restricted analysis is done for only the first term personnel. Random samples are often created for cross validation of regression models.

The processing requirements of tasks begin with the limitations on their initial input. These restrictions and capabilities apply to both incumbent and supervisor data. A task rating may consist of any quantifiable information up to six characters in length. In other words, the possible range would be Ø to 999999. Because of suspected nonlinearity in some scales, it is possible to rescale the task responses. For example, a 1 to 7 scale could be changed to (1, 3, 5, 10, 15, 17, 19).

Not many processing requirements are imposed on background items; however, a few points should be made. Textual information, such as the description an incumbent gives as a job title, is limited to 69 characters in length. The maximum number of characters available for use is 6000. Because of the fatigue factor, rarely, if ever, are there this many variables collected. Whenever a research question arises and the background section of the survey booklet did not solicit a necessary piece of information, other data sources may be searched. A unique identifier is the key to accessing these data bases. It is standard practice in USAF Job Inventories to solicit both name and social security account number (SSAN) in accordance with the Privacy Act of 1974 (PL 93-579). With this information, personnel files may be cross-referenced to obtain test scores, sex or other demographic data. The cost of compliance with the Privacy Act is negligible when compared to the potential benefits.

One area not mentioned thus far is the topic of the data bases (computer files) associated with the CODAP system. They may be categorized into one of two areas: those associated with case/rater data and those containing summary information. The case/rater data files contain information on each individual survey respondent, task titles used in the Job Inventory, and background variable definitions. The first of these, the History file, is created during the Data Preparation and Validation Phase. Another version of this file, created during the clustering process, is known as the KPATH file. Once a summary computation and/or report has been generated, its contents may be captured for later processing and viewing on summary data files. Three types exist within the CODAP system. They are known as the Job Description file, FACSET file and Report file. Both Job Description and FACSET files contain condensed binary summary data. The Job Description has membership definitions and summary data where the FACSET file has only summary data, but its format is designed for fast, efficient processing. The Report file is used to keep a duplicate of a computer report which was produced during the process of an analysis. Because of the refinement of computer systems, it is many times more cost effective to recreate a product as compared to the cost of maintaining copies. It is for this reason the importance of the Report file is quickly fading away. Each of the files described above plays an intricate role in the processing of occupational data. They are the necessary link from which the CODAP package may be considered as a system of interrelated programs. It is important that an occupational analyst know they exist and be informed of their contents. For it is from these files that the data are selected for analysis. Realizing this, a series of computer programs have been developed to index, audit and report information about these files. These programs include DICTXX, JOBSXX, TASKXX, VARSXX, FSINDX, JDINDX, and RPINDX. Descriptions of these programs may be found in the appendix.

APPENDIX A

CODAP Program Summary

Listed below are descriptions of the major computer programs associated with the Air Force CODAP package. A six character computer code along with its title is given. Additionally, the year in which the last major revision occurred and the programmer's name is listed. Finally, a brief summary of the programs capabilities is presented.

AUTOJT Automated Job Typing 1973 - Weissmuller
This program evaluates between-group differences for pairs of Job Descriptions to aid in determining distinct job types within the hierarchical clustering process. Six comparisons are computed and reported for each pair of job descriptions. These evaluations include difference in percent time on each task, percent time spent on each duty, percent members performing each task, number of tasks needed to account for a specified percent of total group time, and average number of tasks performed by each group.

AVALUE Average Values Per Task 1973 - Stacey
AVALUE computes an average value for each task in the Job Inventory. The average value is based on a selected variable for all cases who perform that task. For example, this program might be used to compute the average number of months in service for those members performing each task.

AVGPCT Average by Percent Performing 1972 - Weissmuller
AVGPCT computes an average value for each task in the Job Inventory. The average is based on the percent of members at each level performing. In other words, the average will be adjusted to account for unequal membership within each interval.

CODAPI Interface to Input Data to CODAP 1978 - Weissmuller
The purpose of CODAPI is to provide an interface with other analysis packages. This program will take a COBOL-format file and create CODAP compatible data cards.

CODAPX Interface to Extract Data from CODAP 1974 - Stifle
The purpose of CODAPX is to provide an interface with other analysis packages. This program will take a CODAP History or KPATH file and create a COBOL-format file.

COMGEN Composite Factor Generator 1975 - Weissmuller
COMGEN allows the user to generate a special purpose FORTRAN program to perform operations on vectors from the Job Description file and produce new composite task factors.

CORREG Correlation and Regression Package 1975 - Stacey
This program extracts up to 100 variables from a CODAP History or KPATH file and computes correlation matrices and regression problems. The correlation part computes and prints the correlation matrix, number of

valid cases in the sample, and means and standard deviations of the variables. A series of regression problems may be computed using an iterative technique. The standard and raw score weights for each variable are reported, as well as the regression constants.

CURVES Curve Fitting and Plotting 1975 - Goody/Thew
This program finds the curve of best fit when predicting one variable (Y) from another variable (X) using polynomials. Provisions exist for plotting the curve of best fit with scattergram of actual observations superimposed. At the end of each report a summary is printed which includes means and standard deviations, the correlation matrix, and a regression problem table including RSQ, regression weights and constant.

DIAGRM Diagram of Clustering Process 1975 - Weissmuller
This program generates a treelike diagram that visually displays the order in which groups merged during the hierarchical grouping process. Each node of the tree, representing one stage, displays the number of members at this stage, the KPATH range defining the membership, and the best and average values.

DICTXX Print Variable Dictionary 1973 - Weissmuller
DICTXX will provide the user with a list of variables titles and their respective formats as defined on the history or KPATH file.

DIST2X 2-Way Distribution 1972 - Weissmuller
DIST2X reports a cross-tabulation of values for two variables, either computed or background, for specified cases. In addition to raw frequency, percent of total row, total column, and/or total sample, mean and standard deviations may be displayed. An additional row or column labeled "other" may be added to account for values outside the specified limits.

DUVARS Duty Variable Computations 1972 - Stacey
For each individual case, DUVARS computes the total percent time spent in each duty, the number of tasks performed in each duty, and/or the percent of an individual's responses which are performed in each duty. They may be stored as new background variables on the history or KPATH file.

EXTRCT Reprinting of Reports 1973 - Weissmuller
Extract will reprint any report or group of reports saved on the CODAP Report file during an analysis.

FACCOR Task Factor Correlation 1979 - Thew
This program will extract up to 100 factors on the CODAP FACSET file and compute correlation matrices and regression problems. The computations and reports are like those of CORREG.

FACEXT Factor Extract 1978 - Thew
FACEXT will extract up to 100 vectors on the CODAP FACSET file and write them to a COBOL file. The purpose of this program is to establish an interface to other statistical packages available outside CODAP.

FACGEN Factor Generator 1977 - Thew
The purpose of FACGEN is to modify and/or load task factors for future processing within the CODAP system. Modifying consists of raising values to a specified power, standardizing to a mean of 5.0 and standard deviation of 1.0, or the substitution of rescaled or rank ordered values.

FACPRE Predicted Factors 1979 - Thew
FACPRE will apply the regression equations developed by FACCOR or CORREG. The following items are reported: titles for criterion and predicted factors, the regression equation including titles for the input vectors, the number of observations, product-moment correlation and product-moment correlations squared, and the standard error of the estimate. The mean, moment about the mean, standard deviation, coefficient of variation, and minimum and maximum values are reported in columnar format for easy comparison of the criterion versus the predicted factor.

FACPRT Task Factor Print Program 1977 - Thew
This program allows the user to print any of the factors on the CODAP FACSET file. Its capabilities include calculating and reporting differences, cumulative percentages, categories of tasks, means, standard deviations, and summations. Report formats may be tailored in a variety of ways to meet the needs of different users.

FACSTD Input Standard for Factor Raters 1975 - Weissmuller
FACSTD creates the CODAP Rater History file. Its specifications are similar to that of INPSTD except that raw task responses are stored instead of relative percent time spent.

GRMBRS Group Membership 1973 - Weissmuller
This program produces a detailed report which describes the two groups combining at each stage of the hierarchical grouping process. The information reported includes: stage numbers, number of members in the combined group, KPATH range of the member cases, number of members in each merging group, average overlap between merging groups, and average overlap within the combined group.

GROUP Hierarchical Clustering 1973 - Myer
GROUP is the program that actually performs the hierarchical clustering in the CODAP system. At every stage, the two most similar groups are identified and combined. Once combined, the similarity of this composite group with all groups is reassessed. This collapsing process is continued until only a single group remains. The output from GROUP is used by the KPATH program to incorporate the clustering information back into the mainstream of the CODAP system.

GRPDIF Group Differences 1973 - Weissmuller
This program will report the task-level differences between two job descriptions. Percent time spent or percent members performing each task is used as the basis of computation. Correlations between the percent time spent vectors and between the percent members performing vectors may also be obtained.

That data may be either percent members performing or average percent time spent by all group members.

INPSTD Input Standard for Job Incumbents 1973 - Weissmuller
INPSTD creates the standard CODAP History file. This program stores percent time spent computed from raw relative time responses Duty/Task title cards and History variable definitions are combined with the case data and reorganized in History file format. INPSTD will accept 20,000 cases, 1700 task ratings, 6000 characters of History data per case, and 26 duty categories.

JOBGRP Compute Stage Job Descriptions 1978 - Thew
Given a stage number from the hierarchical clustering process, this program identifies all members in the group formed at that stage and computes a composite job description for those cases.

JOBIND Print Individual Job Description 1973 - Barton
This program prints a job description with specified background information for each individual in a selected group.

JOBPRT Job Description Print Program 1977 - Thew
This program prints job descriptions computed by JOBSPC or JOBGRP. They may be ordered by task, task within duty, or by modules.

JOBSPC Compute Special Job Descriptions 1977 - Thew
Given the membership criteria in terms of computed or background variables, this program identifies all cases meeting these requirements and computes a composite job description for that group.

JOBSXX Audit Job Description 1978 - Weissmuller
JOBSXX is designed to compute and print job descriptions from all cases on a History or KPATH file in three different sort sequences.

KPATH Create KPATH File 1977 - Thew
This program will resequence cases on the History file as defined in the clustering process.

MODSXX Audit Module Definitions 1978 - Thew
MODSXX will print a task listing ordered by module categories. Tasks not included in any module definition will be placed in a module called "Tasks Not Referenced" and printed at the end of the report.

MTXPRT Print Overlap Matrix 1977 - Weissmuller
This program computes the overlap between all pairs of input composite job descriptions and reports these values in matrix form. Overlap may be computed in terms of average percent time spent on tasks or in terms of the number of tasks performed in common.

OVRLAP Overlap of Response Patterns 1975 - Myer
OVRLAP calculates the similarity between all pairs of cases on the History file. The data are arranged into a matrix format for processing by the GROUP program. History files of 7000 cases or less may be input to this

PLOTIT Plot A Task Factor Histogram 1978 - Weissmuller
This program accepts an input task factor and plots a histogram showing the distribution of values.

PLTVAL Plot Mean and Standard Deviation Values 1978 - Thew
PLTVAL is designed to produce a plotting of the mean and standard deviations for a given factor.

PREFAC Predicted Factor Report 1975 - Weissmuller
PREFAC will apply the regression equations developed by TSKCOR or CORREG and produce a task factor representing this predicted factor. This program is being replaced by FACPRE. Input is a job description file rather than a FACSET file.

PROGEN Program Generator 1973 - Weissmuller
PROGEN generates a FORTRAN program from high-level commands and standard FORTRAN statements to perform any operations on the incumbent data found on the History or KPATH file. Its primary use is to add new computed and/or history variables.

PRTVAR Print Variable Values 1974 - Weissmuller
PRTVAR will print the values of selected variables for all cases on the History or KPATH file. The output is often used in the job-typing process.

RANSEL Random Case Selection 1979 - Thew
RANSEL will produce a membership identification vector useable as input to the program SUBSET. Given the membership vector from a current job description, this program will randomly select cases based on a percentage of the total or a specific number

REXALL Inter-rater Reliability 1972 - Weissmuller
This program computes and reports the average inter-rater reliability coefficient of a single rater and the stepped-up reliability coefficient for the total group of raters. The program is used in conjunction with sets of task ratings made by a large number of supervisory personnel. REXALL computes the correlation of each rater's responses to the grand mean vector on those items he or she rated. Other statistics are available to identify deviating raters. The program permits the user to ignore these cases on subsequent passes. Also reported are the mean task ratings and the standard deviations of the ratings for each task.

REXSPC Special Task Factor Computation 1979 - Thew
REXSPC will compute composite task factors, from rater data, based on selected background items.

SETCHK Check Sets of Raw Data Cards 1973 - Barton
SETCHK edits the raw data which will be input to the program INPSTD. Only complete cases are kept for further processing.

SUBSET Create a Subset History/KPATH File 1978 - Weissmuller
SUBSET will create a new History file containing only those members as defined by a composite job description.

TASKXX Duty and Task Title Print 1978 - Thew
TASKXX is designed to print a list of the duty and task titles as entered in the INPSTD program.

TSKCOR Task Factor Correlations 1975 - Weissmuller
TSKCOR is being replaced by FACCOR. Input is a Job Description file instead of a FACSET file.

TSKNDX Task Index 1970 - Stacey
TSKNDX computes and prints the following information for tasks performed by a selected group of members: task titles, mean rating value ("Task Index"), percent members performing, average percent time spent by members performing, average percent time spent by all members, and cumulative sum of average percent time spent by all members.

VARGEN Variable Generator 1973 - Stacey
VARGEN will compute new variables for every case on the History or KPATH file. These variables are based on the individual's task response data. Some of the values computed are: average task difficulty per unit time spent (ATDPUTS), overlap of an individual's time spent with a given job description, and time spent over a specified set of tasks. When comparing a rater with a given task factor policy, VARGEN can compute the sum of absolute difference, sum of squared differences, and Pearson roduct-movement correlations.

VARSUM Variable Summary 1973 - Weissmuller
VARSUM computes and prints frequency distributions for specified intervals, reports total frequency counts, and calculates means and standard deviations on selected background and computed variables for any group of individuals whose job description has been generated by JOBSPC or JOBGRP.

VARPCT Variable Percent Summary 1973 - Weissmuller
VARPCT produces the same output as VARSUM except that instead of reporting frequency counts, percentages are used.

VARSXX Variable Value Audit Distributions 1978 - Thew
VARSXX is designed to produce a listing of the actual responses, either history or task, given by all incumbents on a History or KPATH file. This output is very useful for establishing specifications for the VARSUM and JOBSAC programs. It is also used as an auditing tool of the History file after INPSTD.

REFERENCES

Morsh, J. E., & Archer, W. B. Procedural Guide for Conducting Occupational Analysis in the United States Air Force. PRL-TR-67-11, AD-664 037. Lackland Air Force Base TX: Personnel Research Division, Sep 1967.

Weissmuller, J. J., & Kauffman, B. Data Processing Problems in Collecting Job Survey Information. Paper presented at 18th Annual Conference of the Military Testing Association, U. S. Navy, Gulf Shores AL, 18 Oct - 22 Oct 1976.

Weissmuller, J. J., Barton, B. B., & Rogers, C. R. CODAP: Programmer Notes for the Subroutine Library on the Univac 1108. AFHRL-TR-74-85, AD-A004 086. Lackland Air Force Base TX: Computational Sciences Division, Human Resources Laboratory, Oct 1974.

Stacey, W. B., Weissmuller, J. J., Barton, B. B., & Rogers, C. R. CODAP: Control Card Specifications for the Univac 1108. AFHRL-TR-74-84, AD-A004 086. Lackland Air Force Base TX: Computational Sciences Division, Human Resources Laboratory, Oct 1974.

McFarland, B. P. Job Analysis of the Medical Service Career Field. AFHRL-TR-73-36, AD-775 720. Lackland AFB TX: Occupational Research Division, Human Resources Laboratory, Oct 1974.

Ward, J. H. Jr., Hierarchical Grouping to Maximize Payoff. WADD-TN-61-29, AD-261 750. Lackland Air Force Base TX: Personnel Laboratory, July 1963.

Christal, R. E. The United States Air Force Occupational Research Project. AFHRL-TR-73-75, AD-774 574. Lackland Air Force Base TX: Occupational Research Division, Human Resources Laboratory, Jan 1974.

McFarland, B. P. Potential Uses of Occupational Analysis Data by Air Force Management Engineering Teams. AFHRL-TR-74-54, AD-A000 047. Lackland Air Force Base TX: Occupational Research Division, Human Resources Laboratory, July 1974.

CODAP: MULTIPLE CLUSTERING APPLICATIONS

Johnny J. Weissmuller

Air Force Human Resources Laboratory
San Antonio, Texas 78235

INTRODUCTION

Within the United States Air Force, as well as many other military and civilian agencies, there is a fundamental technology which supports both the operational and research occupational analysis programs. The core of this technology is called CODAP, an acronym for the Comprehensive Occupational Data Analysis Programs. The CODAP "system" is a set of analysis tools and procedures which use, as raw material, information provided by the members of the occupational field being studied. Although it composes only 15% of the system, the hierarchical clustering procedure is the hallmark of the CODAP system. These clustering programs and analysis guidelines give CODAP the highly desirable ability to identify occupational structures and substructures based on the similarities of time spent reported by job incumbents. Through proper analysis, these structures are broken down to identify clusters of related jobs which are, in turn, broken down to reveal distinct job types. These clusters and job types may be used to revise classification systems, assess job related skills, verify the relevance of training programs and a host of other applications in which an accurate knowledge of job content at the task level is desirable.

This paper traces the development of analysis techniques which will facilitate routine operational studies as well as research studies involving multiple clustering applications. Three research applications are discussed. For each application, a statement of the basic problem is presented followed by the general findings in the study. After this overview, the details of the study are analyzed to highlight the relevant procedures. These applications were chosen because each used the hierarchical clustering procedures twice in the course of its analysis and attempts were made to compare the two generated clustering solutions. The insights gained in these applications contributed directly to the development of the Cross-KPATHing Technique and the Homomorphic KPATHing Technique explained in the second half of this presentation.

The motivations for administering multiple surveys varied, but the same questions arose in each case: "How does one compare two different clusterings and what conclusions can be drawn?" The initial approach taken was to determine if the same individuals formed groups under both clusterings. In other words, if a given set of twenty people formed a group identified on clustering diagram 1, did they also form a group that was identifiable on clustering diagram 2? This approach is not a general solution since it is based on the premise that a common subset of individuals can be found in both clusterings. There were a sufficient number of members in common to use this approach in all three applications discussed. The techniques developed for studying multiple clustering applications are limited to instances in which this condition is met, while the techniques for use in single clustering applications are completely general.

PART 1: THE APPLICATIONS

To present the applications in the proper perspective, a few key points should be reviewed. Basically, the hierarchical clustering procedure is a mathematical method for identifying homogeneous groups of people based on some measure of their similarity. Currently, there are only two measures of similarity available in the CODAP system. The standard option evaluates the percent of total time spent in common between all pairs of survey respondents. The secondary option will evaluate the percent of items checked by both respondents as a function of the number of items checked by each individually. The clusterings performed in these applications used either the standard option, referred to as a "TIME" matrix, or the secondary option, referred to as a "TASK" matrix. In about two years the Profile Analysis extension to CODAP will be released and will contain 20 additional measures of similarity. With this in mind, the question of addressing the relationships between multiple clusterings takes on even greater importance.

Each job incumbent surveyed receives a Job Inventory Booklet which is composed of a background section and a task inventory section. The similarity measures are computed solely from the data in the task inventory section. The three applications discussed in this paper may be further characterized by the content or timing of their respective "task inventories." The first application, chronologically, was performed by the United States Air Force Occupational Measurement Center (OMC) and it compared a routine time-rated task inventory to a specially prepared checklist of electronic principles. The second application, performed by a Canadian utility, compared a routine time-rated task inventory to a time-rated inventory composed exclusively of tools and equipment. The third application was done by the Air Force Human Resources Laboratory (AFHRL), and it compared a single time-rated task inventory to itself by surveying the same job incumbents twice, eighteen months apart.

APPLICATION 1 (Ruck, 1977)

The first of the three applications was the Occupational Measurement Center's Electronic Principles Inventory (EPI). OMC had been tasked by a higher headquarters to develop, administer and analyze a survey composed of electronic principles. The data from this survey were processed with the CODAP system and the resulting information was used to evaluate fundamental training courses across all electronics-related career fields. Fifty-nine career ladders were surveyed and reviewed in this project. About 63,000 personnel are directly involved in various electronic fields in the United States Air Force and approximately one-half million dollars is spent each day for electronic principles training. For this reason the results of this study are still being evaluated.

Though not in the original plan, the data were run through the clustering programs to assess the potential impact of recommendations to merge or subdivide several electronics career ladders. This led to the question: "How do the groups identified using the EPI data compare the groups identified using the routine time spent data?" As reported in "Job Knowledge Analysis: A New Approach" at the 1976 Military Testing Association (MTA) Annual Conference, the membership overlap between corresponding clusters from the normal time spent analysis and the new Electronic Principles Inventory checklist was 80%. For the corresponding job types, the average membership overlap was 73%. It was concluded that both survey instruments were capturing the same structure and that one might accurately predict the job an individual performs, given that one knows the knowledges that the individual uses on the job.

The procedure used to arrive at the conclusion that the two surveys were capturing the same job structure needs to be examined in more detail. The basic comparison procedure used in all three applications is as follows: perform two clusterings, perform two job-typing analyses, identify corresponding groups on the two diagrams, and finally, compute the membership overlaps. Because the computer programs limit the clustering analysis to job incumbents surveyed with the same task list, the time spent data was automatically restricted to a single career ladder. To establish a comparable EPI sample, booklets from incumbents in the selected career ladder were extracted from the pool of EPI respondents which covered all fifty-nine ladders. The two clusterings were performed. The first was actually accomplished as part of the routine OMC occupational survey program using the time spent data and the standard similarity option evaluating the percent of time spent in common. The second clustering was done eight months later as part of the EPI project and used the secondary option to compute similarity based on the percent of items checked in common.

Each clustering was job-typed; that is, analyzed to identify major clusters of related jobs and distinct subcomponents called job types. Job-typing is currently closer to an art form than an exact science. The computer programs produce a mathematically-defined structure depicting similarity based on time spent, but it is up to the occupational analyst to decide which groups in the hierarchical structure represent essentially different jobs. Two points need to be made. First, superficial differences may appear in the solutions from independent analysts but are insignificant from a management decision-making perspective (Watson, 1973). Secondly, job-typing is always accomplished in the context of some overall purpose which affects the decisions being made. A mathematically significant difference between two groups based on a totally insignificant task does not, in itself, provide grounds for reporting the groups as "essentially different." The significance of a task can only be assessed in the context of this overall purpose. For example, if the two groups differed greatly on the task "Make photocopies" but showed no difference on the task "Prepare agency budgets," would the groups be reported separately or combined into a single job? In a typical analysis which attempts to identify differences based on aptitude or training requirements, the groups would be combined because making photocopies is insignificant in that regard. In a job redesign analysis, however, it may be very significant that low-level tasks are being performed by high-level personnel and the groups may well be reported separately to highlight a situation requiring further attention. In this OMC application, both the time spent and EPI data were job-typed to differentiate groups based on aptitude and training requirements.

After the job-typing had been completed, the question of membership overlap was addressed. Because the two samples did not contain exactly the same individuals, the first step was to identify and remove the individuals in each sample who did not appear in the other sample. After the clusterings had been resolved down to a common subset, the next step was to tentatively identify the corresponding groups on the two diagrams. At that point in time there was no method to aid in establishing these correspondences other than a visual inspection of the background data to detect a large number of people in common. The computations were accomplished by going to one diagram, selecting a cluster and comparing its members to the members in the corresponding group. The proportion of common membership was computed by counting the number of people who appeared in both of the corresponding groups and dividing that number by the count of people in the selected cluster. Multiplying that proportion times 100 yielded the percent of membership from the selected cluster in common with its corresponding group. This process was repeated for all clusters and job types, first mapping the "time spent" groups into the EPI groups and then vice versa. Averaging the resultant percentages for the clusters and job types produced the overall figures of 80% and 73% respectively.

The key question, both figuratively and literally, was: "How were the individuals identified as being the same person in both clusterings?" It is standard practice in USAF Job Inventories to solicit both name and social security account number (SSAN). In this project, individuals were tracked manually by name--a labor intensive process. Had personal identifiers not been routinely collected, this cross-matching problem would have been a major stumbling block in the analysis. The original time spent survey was accomplished in the standard occupational survey program with no special considerations for future cross-referencing needs. Only after the EPI survey was administered eight months later did it become apparent that a matching requirement would arise. Because of the statistical nature of the CODAP system, the cost of compliance with the Privacy Act of 1974 (PL 93-579) is negligible when compared to the potential benefits. For example, social security account number (SSAN) is routinely used to cross-reference into personnel files to obtain test scores, sex or other demographic data for further statistical breakouts whenever a research question arises and the background section of the survey booklet did not solicit a necessary piece of information. Without the SSAN the alternatives would be to leave the research questions unanswered or resurvey the job incumbents--a very expensive alternative. The name, rather than the SSAN, was used in this OMC study because the matching was done by hand. Had an automated approach been available to cross-match information about the structure, it would have relied on the more stable identifier, social security account number, as the key.

APPLICATION 2

The second application was done by one of the world's largest utilities, which produces electrical power using hydroelectric, coal, gas, and nuclear generators. They were interested in reevaluating their training programs for maintenance personnel. It was clear that the type of generating system maintained would vary from plant to plant, but there was apparent commonality in the tools and maintenance equipment used. The existing training program revolved around systems maintained, with very little attention given to developing proficiency in using basic maintenance equipment. The question to be answered was: "Is there sufficient commonality in the maintenance tools used across different systems to justify independent training programs for tools in addition to systems maintained?" While in the first application two clusterings captured the same structure, in this case, the clusterings identified widely variant structures, lending some support to the argument for separate training.

The primary research scientist on this project prepared a single survey booklet containing a dual "task" inventory. The first was a conventional task inventory covering the maintenance of all types of energy-producing systems. The second inventory was composed exclusively of tools and equipment used in maintenance work. This inventory presented the tools and equipment subdivided into various categories such as "hand tools," "safety equipment," etc. The incumbents were asked to time rate this tool list as if it were a normal task inventory. The ratings on the inventories were independent; that is, a task was never rated relative to a tool nor vice versa.

When the survey booklets were returned from the field, they were optically scanned and entered into the CODAP system twice. The first time the "system maintained" tasks were declared the task statements for purposes of computing time spent and the "tool" data were declared background information. The second time, the "tool" tasks were used for time spent calculations and the "system maintained" tasks became the background information. Each set of data was run through the clustering programs using the standard percent of total time option.

After he had job-typed both samples, the research scientist consulted with Mr Hendrick Ruck of AFHRL who had worked on the EPI project at OMC. Mr Ruck outlined the labor intensive steps ahead. A careful review of the clustering printouts revealed only a single corresponding job type--that of the welders who had long ago established themselves as an independent occupational entity. In that job type, 18 members were found in common between the 20- and 22-member groups depicted on the two diagrams. As far as the other job types were concerned, it seemed hopeless to identify even tentative correspondences. The people who were clustered based on "systems maintained" broke cleanly by type of generating plant, while those same people, when clustered on "tools used," also had fairly clear breaks along the lines of logical tool families. There were few easily identifiable membership correspondences between groups identified in the two clusterings. The tool clusters, with the exception of welders, could not be identified with any of the existing training courses, which, on the other hand, mapped quite well into the "systems maintained" diagram. Because of the labor intensive process required, no exact numerical evaluation of common membership was attempted between the widely variant structures.

APPLICATION 3

The third application was performed by the Air Force Human Resources

Laboratory (AFHRL) as part of a job satisfaction research project. The basic question to be answered was: "To what degree does a change in job content affect job satisfaction?" One of the classic problems in this type of study is: "How does one change the job content for a subject without introducing the Hawthorne effect; that is, changes in job attitude unintentionally induced by the observer?" The experimental design of this study employed an elegant solution to this problem; namely, assess both job satisfaction and the job content for a large number of people, wait a reasonable length of time to permit job content to change naturally (reassignments, widened responsibilities, promotions) and resurvey the same people with the same inventory with perhaps a few additional job satisfaction items. This study demonstrated that knowledge of job content at two points in time can greatly enhance the accuracy of job satisfaction predictions.

Periodically the question is raised: "How has a particular career field changed over time?" The question is usually addressed by resurveying the career field incumbents, job-typing the new respondents, and drawing some conclusions regarding the nature of the differences between job types existing at time 1 versus those existing at time 2. This type of analysis is known as a Time 1 - Time 2 study. This AFHRL research stream falls into that general category with two additional distinctions. First, because this application needed to assess job satisfaction for the same individual at two points in time, it could include only those individuals who were present for both administrations. Restricting the sample in this manner resulted in some logical discrepancies between the Time 1 and Time 2 clusterings. For example, low-level apprentice jobs which existed at Time 1 had no counterpart at Time 2 and what appeared to be a homogeneous group of supervisors at Time 1 split out at Time 2 to become a group of first-line (working) supervisors and another group of second-line (administrative) supervisors. Because the inventories had been administered eighteen months apart, the entire workforce had gained more experience and the job difficulty index was higher at Time 2 than at Time 1 for all corresponding job types.

The second characteristic which distinguished this analysis from the typical Time 1 - Time 2 study was the fact that an identical task list had been used in both administrations. A follow-up Time 2 administration is usually 3-4 years after the first survey and the task list is updated to match the changes in the career field. This study was predicated on the assumption that in eighteen months individuals would change jobs, but that the job types themselves would remain relatively stable. The data support this assumption. Since this study used identical task inventories, precise Time 1 - Time 2 job type correspondences could be established through automated means. By entering the job descriptions for all job types identified at Time 1 and Time 2 into a CODAP program, a matrix was produced which reported the percent time spent in common for all possible pairs of groups. Those Time 1 - Time 2 combinations having greater than 70% time in common could be declared corresponding groups. The 70% time may seem low but it should be recalled that all job types at Time 2 had shown a higher job difficulty index because of the effects of widened responsibilities and this was reflected in the distribution of time spent. As a point of reference, however, the 70% time overlap roughly translated into correlations in the .90's between the time spent vectors in matched groups. Correlations for the same vector in nonmatching groups ran in the .60's and below. In other words, job content at the task level was used to establish the job type correspondences and this meant that individuals could transfer from one job to another without causing a drop in the stability measure by which job types were identified across time.

As the aim of this project was to measure the impact of change on job satisfaction, a reliable indicator was needed to identify those individuals who had experienced a substantial change in job content. A subgoal was to find a way in which to categorize, if not detail, the nature of that change. It was decided that job type membership constituted a categorization of job content and that membership in non-corresponding job types at Time 1 and Time 2 represented both a substantial change and an indication of the nature of that change. A method was needed to identify each person's job type at Time 1 and Time 2 on a single file to permit this information to be used in the prediction of job satisfaction. A new technique, tentatively called "Cross-KPATHing," was under development and was used to accomplish this task. This technique detailed the movement of people from one job to another so effectively that the primary researcher on the project, Mr Kenn Finstuen, was able to present this information graphically in his technical report now in press. These Time 1 - Time 2 personnel movements will be reported as the "Dynamic Migration Chart."

As a by-product, the Cross-KPATHing technique provided a potential solution to the related problems of the two previous applications and led to further insights which may benefit future operational analyses. The second part of this paper is devoted to a general discussion of the KPATH sequence number, the Cross-KPATHing Technique, the Homomorphic KPATHing Technique and their relationship to these and other applications.

PART 2: The KPATH Sequence and Its Usages

The term "KPATH" is borrowed from the branch of mathematics called topology, where it is generally used to indicate some order in which the nodes of a network are traversed. In the CODAP system the term is used to denote an order in which to arrange the job incumbent data to facilitate analysis. When the data for individual incumbents are reported in this order, it aids in the identification and analysis of job types by accentuating the natural breaks between groups. This KPATH sequence is also the basis on which the cluster merger diagrams are organized and printed.

The KPATH sequence is developed as part of the hierarchical clustering process and has the essential property that all job incumbents in any cluster or job type may be identified by a single unique range of KPATH sequence numbers (Phalen, 1975). A KPATH sequence number is the position or rank order assigned to a job incumbent in the process of arranging all cases in KPATH sequence. Although the KPATH sequence number is more an attribute of the structure than of the job incumbent, it is routine practice to ascribe KPATH sequence numbers to individual cases and actually make that number part of the record containing the person's responses. When this KPATH sequence number becomes part of the incumbent's data record, one may identify cluster or job type groups in the same way that one identifies groups based on grade level, months on the job or current job classification. Within the CODAP system, this also permits job type information to be used with any other data to produce hybrid groups such as those people in a given job type who have 1 to 24 months on the job.

The Cross-KPATHing Technique

For purposes of tracking cases across multiple clusterings, the procedure of ascribing a structural attribute to a specific individual is at the heart of the solution. By allowing individuals to "carry over" their KPATH sequence number from one clustering to another, a cross-referencing system is established between the structures. Each additional clustering generates a new KPATH sequence and assigns a new KPATH sequence number to each individual. For clarity, let the KPATH sequence numbers for each case be identified as KPATH1, KPATH2, KPATH3, etc, where the 1, 2, 3 represent the clustering which generated that particular sequence. After the second clustering then, each job incumbent has two KPATH sequence numbers, one (KPATH1) which can be used to identify membership in job types from the first clustering, and the other (KPATH2) which can be used to define membership in the job types from the second clustering.

There is a program in the CODAP system which will print a two-way distribution (cross-tabulation) reporting the number of people who fall into specified ranges of values for any two variables. One could report, for example, a frequency distribution table in which the rows identified grade levels, say E-1 to E-3, E-4 by itself, E-5 to E-7 and E-8 to E-9. In the same table, the columns may identify months on the current job and may be broken down into various ranges such as 0 to 12 months, 13 to 24 months, 24 to 48 months, and 49 to 360 months. Once the new KPATH1 and KPATH2 variables have been moved to a single data file, they can be used in this program to report the membership composition of any job type in terms of the number of people which came from each of the job types in the other clustering. By reporting row and column percentages, one can define membership composition in both directions in a single table. This procedure of moving multiple KPATH sequence numbers to a single data file and comparing clusterings using two-way distributions is called the Cross-KPATHing Technique.

One of the problems not mentioned in the OMC analysis of "time spent" versus "knowledges used" was that the job-typing in the first clustering identified 12 job types while the second job-typing identified only 11. The overall 73% overlap figure was based on the fact that an EPI group not identified as a significant group in the job typing process had been located and associated with the extra "time spent" job type. The Cross-KPATHing Technique does not require a completed job-type analysis in order to function. In fact, instead of being predicated on the results of a job-typing, this technique may actually facilitate and guide the job-typing process.

This possible procedure begins after the two clusterings have been performed by moving the two KPATH sequence numbers for each case to a single data file. The cluster merger diagrams are then used to identify the smallest, non-intersecting groups with five to eight members in each clustering. This process is repeated for group sizes 10, 20, 30, and so on. The two-way distribution program can then test and report all these proposed job-typing configurations in a single computer run. The configuration which provides the highest membership overlaps may be a combination of levels; that is, some groups matching best at the 5-member group level, some at the 20 and still others at the 50-member group level. This procedure does not replace the job-typing process, but simply provides a starting point with a little more information than would have otherwise been available.

In some cases the level at which the optimal membership overlap occurs may provide insights. In the example which compared "systems maintained" versus "tools used," reporting out very small groups might have strengthened the case for separate training. For example, if every group eventually called a job type in the "systems maintained" clustering was found to contain distinct subgroups representing all of the job types from the "tools used" clustering, and vice versa, it would have indicated that the two structures were uncorrelated. Under these conditions, the argument for training programs based on either structure should be equally strong.

When used in the AFHRL Time 1 - Time 2 study, the Cross-KPATHing Technique was still under development. The original method of reporting common memberships used the CODAP variable summary program, but it required that job descriptions be computed in advance and it could report percentages in only one direction. The other problem was that the processing affected the appearance of the second diagram and that was considered unacceptable.

The current Cross-KPATHing Technique, however, has overcome these problems and is being used in an AFHRL research study investigating job incumbent interpretations of the points on a rating scale. To date, six scale transformations have been tested and run through the clustering programs. All these clusterings are being tracked on a single data file containing six unique KPATH sequence numbers and the structures are being compared using the two-way distribution program. Other AFHRL research projects will study the clustering of tasks based on various similarity measures and the Cross-KPATHing Technique will also be used in those applications.

The Homomorphic KPATHing Technique

The objection to the original Cross-KPATHing Technique was that it reordered the case data prior to clustering and it is a long-recognized fact that the input order of the cases has a direct impact on the KPATH sequence numbers. Because of the relationship between KPATH sequence numbers and the cluster merger diagram, a change in input sequence can produce a second diagram from the same data which has a different visual configuration and perceptual impact. In both cases, however, the mathematical relationships between individuals, groups, and stages would remain unchanged. The original method of Cross-KPATHing was considered unacceptable, because it introduced a psychological distraction into the already subjective job-typing process. Two approaches to solving that problem were explored. The first, changing the Cross-KPATHing technique to avoid reordering the input cases, proved to be the faster approach. The second, standardizing the assignment of KPATH sequence numbers, regardless of input sequence, is the subject of the remainder of this paper.

The hierarchical clustering procedure is an iterative process (Ward, 1961; Archer, 1966). At each iteration, called a stage, a new composite group is formed from two individuals and/or previously formed groups. Upon completion of the clustering, the iterative results are displayed pictorially in a report called a cluster merger diagram (Phalen, 1973). This cluster diagram prints five items of information for each reported stage. The item of relevance to this discussion is the range of KPATH sequence numbers which define the membership in the group formed at each stage. The KPATH range is the controlling factor in printing the diagram. The stages are arranged from left to right to insure ascending values for the KPATH sequence numbers across the topmost level of the diagram. In general, there is no practical significance to the position of a stage with respect to the left or right sides of the diagram except to indicate, in some sense, a relationship to the first case input to the clustering. Analysis guidelines point out that it is desirable to select several booklets filled out by senior managerial personnel and recode their booklet numbers to insure that they are processed first by the CODAP system. Several booklets should be recoded to insure that at least one will pass all the initial processing and acceptability checks and become the first case actually entered into the system. This procedure will tend to place managerial and supervisory personnel to the left on the diagram and the journeymen and apprentices to the right.

The KPATH sequence is currently generated during the clustering process by noting the order in which members are added to a group. An ordered list of members is associated with each group. The members are uniquely identified by the sequence in which they were originally entered into the clustering programs. Thus, when two groups are merged, at any stage, the membership lists from both groups are also combined to reflect a single, larger group. The method by which these membership lists are combined is the primary point of interest. The current decision rule for combining the two lists states that the list containing the case with the lowest input sequence number will be used first. Following the last case in that first list, the list from the second group will be appended. In other words, even when combining two individuals, the resulting membership list will contain the case having the lowest input sequence number followed by the other case. This use of the case input sequence number to order the combination of the membership lists creates a KPATH sequence that is arbitrary and that is the core of the problem.

If the combining of the membership lists is based on an attribute of the mathematical structure, it may be possible to generate a unique KPATH sequence that is a function of the structure to the greatest degree possible. There are four items of information reported for each stage on a diagram which may be used for this ordering procedure. These items include stage number, number of members, the similarity measure which caused the two groups to combine ("Best" or "Between" value), and the homogeneity measure for the newly formed group ("Average" or "Within" value). By selecting one of these four variables, a decision rule may be established to order the combining of membership lists based on the relative size of those values found in each of the two merging groups. With the apparent exception of stage number, all the variables identified are subject to the possibility of tied-values. Because the clustering actually begins with single-case "groups", stage number and previous "between" values are not defined and the group size and "within" values result in a large number of ties whenever individual cases are merging.

Numerous alternatives have been considered in the search for a tie-breaking decision rule that is consistent with, if not directly related to, a structural attribute. Because the decision rule must apply at the level of individual cases, an initial reaction is to order cases on a job-related background characteristic such as grade, months on the job, or months in the career field. A better variable for resolving ties might be the average number of tasks performed by an individual or group. Because the CODAP system automatically computes a variable for each case indicating the number of tasks performed, the question of missing data is removed. In general, however, variables are subject to ties and, in addition, are afflicted with the problem of missing data. For this reason, the most appropriate decision rule is actually an ordered series of subrules, culminating in some arbitrary, but definitive rule which may be required as an ultimate tie-breaker.

If one is interested in generating an invariant KPATH sequence, the following decision rule will produce an almost invariant solution. Although this rule does reflect some preferences, it is being presented solely as an example, not as a strong recommendation. Let the primary ordering be based on group size. If two groups of unequal size are merging, place the membership list of the larger first. In case of a tie in group size and the size is greater than one, place the membership list of the group which was formed at the higher stage number first. This condition is the same as placing the group with the highest "between" value first, except that stage number is unique and that resolves the question of ties. In the case of a tie in group size and the size is equal to one, that is, two individuals merging, place the individual with the highest number of task responses first. In case of a tie on that criterion, arrange the cases based on input sequence number. The last condition is included to provide a definitive condition, although it is strongly hoped that it would never need to be invoked. Applying the rule does not provide a totally invariant sequence, but it is very close and permits reversals of individual cases only within the context of a single pairing. This rule should result in a KPATH sequence which greatly accentuates the breaks between groups in the ordered data prints. These observations are based on the fact that all accretions to existing groups would be attached to the end of the current membership list and the perceptual impact would be that of a large, stationary mass, slowly attracting and absorbing its smaller neighbors to the right on the diagram.

The foregoing discussion was predicated on the assumption that one wished to establish an invariant KPATH sequence number to overcome objections to the Cross-KPATHing Technique. It is not at all clear that the use of an invariant KPATH sequence is desirable. It was suggested, in part, because of the possible impact on the subjective process of job-typing. Rather than making the sequence invariant under all conditions, perhaps it should be tailored to meet the needs of a given analysis. Recall that job-typing is always accomplished in the context of some overall purpose and that this purpose is expected to affect the decisions made. With this in mind, it is somewhat ironic that the Homomorphic KPATHing Technique will probably realize its greatest potential, not by always producing diagrams of the same shape, but by molding the same structure into multiple shapes to meet the needs of changing applications. The Homomorphic KPATHing Technique is simply the procedure of defining the decision rule by which these membership lists are combined. An attempt should be made to identify a criterion which demonstrates a great range of values all the way down to the case level and hence avoid ties and the need for additional rules.

If one is no longer attempting to capture an invariant KPATH sequence, the requirement that the decision rule be primarily based on an attribute of the structure may be removed. Under these relaxed constraints, the search needs to start anew, looking for a basis on which to order cases, jobs, and eventually diagrams. The recurring request from operational occupational analysts has been to organize the diagram from left to right, placing more demanding jobs to the left and less demanding jobs to the right. Because the jobs themselves are composed of time-rated task vectors, the existence of any task vector defining the relative "importance" of the tasks will permit such an ordering to take place using the Homomorphic KPATHing Technique. After a traditional clustering has been accomplished, this "importance" vector may be applied to each job to compute a time-based weighted composite using a existing CODAP program. If the "importance" concept is actually a hybrid of the task-based composite and additional information, a second CODAP program may be required. For example, the job difficulty index is just such a hybrid, using not only the average task learning difficulty per unit time spent, but also the number of tasks performed. When that composite is used as the primary ordering value, ties are rare due to the seven or so digits of precision and the diagram will come out in the desired high-to-low job difficulty sequence.

It is important to note that the introduction of a second factor does not in any way affect the mathematical structure being defined by the time spent ratings. This factor is being introduced only to order the sequence of presentation of those groups (jobs) defined by the time spent structure. What was an arbitrary decision rule is being replaced with a decision rule suited to the needs of the application at hand. The job difficulty index was but one example. For another, Mr. J. Britt Kauffman (Kauffman, 1978) proposed a task-based point system for private sector applications that would order jobs based on recommended compensation. In any case, the introduction of an ordering principle will ease the job of the occupational analyst by aiding in the job-typing process and by providing a consistent theme on which to transition between groups in the occupational survey report.

REFERENCES

Archer, W. B. Computation of Group Job Descriptions from Occupational Survey Data. PRL-TR-66-12, AD-653 543. Lackland AFB TX: Personnel Research Laboratory, Aerospace Medical Division, December 1966.

Kauffman, J. B. A Task-Based Point Method of Job Evaluation. Unpublished Masters Thesis, University of Texas at Austin, 1978.

Phalen, W. J., & Christal, R. E. Comprehensive Occupational Data Analysis Programs (CODAP) Group Membership (GRMBRS/GRPMBR) and Automated Diagramming (DIAGRM) Programs. AFHRL-TR-73-5, AD-767 199, Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, April 1973.

Phalen, W. J. Comprehensive Occupational Data Analysis Programs (CODAP): Ordering of Hierarchically Grouped Case Data (KPATH) and Print KPATH (PRKPTH) Programs. AFHRL-TR-75-32, AD-A016 724. Lackland AFB TX: Occupational and Manpower Research Division, August 1975.

Ruck, H. W. (Ed) The Development and Application of the Electronic Principles Job Inventory. USAFOMC-TN-77-02. Lackland AFB TX: USAF Occupational Measurement Center, Air Training Command, October 1977.

Ward, J. H., Jr. Hierarchical Grouping to Maximize Payoff. WADD-TN-61-29, AD-261 750. Lackland AFB TX: Personnel Laboratory, Wright Air Development Division, March 1961.

Watson, W. J. The similarity of Job Types Reported from Two Independent Analyses of Occupational Data. AFHRL-TR-73-58, AD-776 777. Lackland AFB TX: Occupational Research Division, February 1974.

OCCUPATIONAL ANALYSIS IN THE CANADIAN FORCES

Cdr Fred J. Hawrysh

LCol Carl A. Leech

CANADIAN FORCES DIRECTORATE OF MILITARY OCCUPATIONAL STRUCTURES

BACKGROUND

1. Occupational Analysis in the Canadian Forces (CF) is the responsibility of the Directorate of Military Occupational Structures.

2. DMOS is responsible to develop and control the Military Occupational Structure (MOS) for all the Canadian Forces including the Regualr Force, the Reserve Force and any Emergency/Mobilization Forces required in the event of crisis or war.

3. The terms of reference of DMOS include:

a. Planning and developing the MOS for both Officers and Other Ranks and making adjustments to the MOS as required, including the addition or deletion of classifications or trades;

b. Validating proposals for new or special personnel qualification requirements as a result of changes in roles, organizations, new equipment or new technology;

c. Determining the impact of capital equipment acquisitions on the MOS;

d. Conducting Occupational Analysis of functions, classifications, trades and special problem areas as required; and

e. Developing, issuing and controlling the personnel specification for each classification trade or specialty.

4. The CF MOS is defined as the framework within which all personnel are recruited, selected, trained, posted, promoted, paid and employed. It is used to provide a means of identifying the personnel skills and knowledge necessary to perform the jobs which have to be done by the members of the CF.

5. The MOS is essentially a grouping of occupations:

a. for officers into:

(1) Classifications - the basic grouping to which an officer is assigned and within which his ca-eer is monitored and guided regardless of his employment;

(2) Sub-classifications - a group within a classification who require additional skills and knowledge to perform duties which form a significant part of the overall function of the classification; and

(3) Specialties - additional special skills and knowledge required to perform a specific job.

b. Other Ranks (enlisted) are broken down in a similar fashion into trades and specialties.

6. Each classification, sub-classification, trade or specialty is described by a specification which defines the scope of employment, progression in the career field, medical and security requirements, working conditions and specific hazards, task listings, trade qualification levels, task involvement, and knowledge and skill levels.

7. The specification is based on Occupational Analysis data and recommendations. It is a management policy and control document used throughout the CF personnel management system. This includes establishing recruiting and selection standards, the classification trade and rank structures, personnel development and career progression programmes, manpower utilization planning, personnel and unit establishment requirements, performance evaluation, establishment of compensation and benefits packages, and, probably the most important of all, establishing training and training validation requirements.

INTRODUCTION

8. The Canadian Armed Forces have been unified for more than 10 years. Unification of the Forces took the three separate services of Navy, Army and Air Force and combined them into one force with one uniform, one rank structure etc. The aim was to reduce personnel and training costs and direct more of the defence budget toward new equipment. The whole exercise was carried out with some haste in order to show immediate cost savings. For example, some 350 enlisted trades of the three services were reduced to less than 100 for the new force. Schools, courses, and training functions were combined, reorganized, reduced or closed. We went from nearly 130,000 all ranks uniformed personnel down to about 80,000. A new Military Structure was introduced and with its introduction came the deletion, dilution and/or combination of some areas of expertise which in retrospect should not have been touched. Whatever mistakes were made did not show up or cause undue hardship immediately because the experienced people were still available and were used as required even if their employment didn't fit the mold of the new occupational structure.

9. At the same time of course, having abolished the three services, their headquarters, staffs and schools, we had to develop a new training system - appropriately enough we called it the Systems Approach to Training (SAT). The SAT was to train to a Course Training Standard (CTS) based on the requirements of the Personnel Specification which in turn was to be based on Occupational Analysis. The system stressed training only to the Minimum Standard necessary to do the specified job at the specific rank and trade level. In a very short time we went from a series of rather luxurious individual training systems to a combined minimum standard system within a new occupational structure and personnel management system.

10. Now some 10 to 12 years after the whole thing started we've finally begun to face up to our many problems. Our political bosses are supplying us with many items of new equipment, including tanks, armoured vehicles, missiles, ADP systems, Long Range Patrol Aircraft and promise of new patrol frigates and fighter aircraft. Most of these items are solutions to our personnel structure and training system. Of course, there are other problems demanding solutions such as:

(a) that pool of widely trained and experienced tradesmen from our old single service days is rapidly disappearing through retirement;

(b) the minimum standard training system is not producing a satisfactory technician/tradesman; and

(c) the training system needs a method to keep up with the new technology and equipment being introduced.

11. Most importantly it appears that the MOS did not grow and change with the system and times and now may be one of our largest problems. Since it is t start point in the Personnel Management System it affects all the other parts including trade and rank structures, training standards and career management. The Structure and Personnel Standards are in urgent need of re-examination and update to cater to the new technology and introduction of women into a much wider scope of employment.

12. A very large portion of our personnel budget and indeed total forces budget, is used for training and the training budget is so tight that we must be very careful to avoid duplication of training or over training. Since the training standards are based on the personnel specifications, any change to the structure or specification must be done with due consideration to training. At the same time it is essential that we provide the Commanders with personnel whose training will enable them to perform the assigned mission.

13. We believe that our traditional methods of conducting an Occupational Analysis with its emphasis almost totally on tasks currently being performed doesn't provide the detailed information needed to develop and revise the occupational structure, produce personnel and training standards which will be valid for our future equipments and tasks. We have developed several innovations to our occupational surveys and the way we go about doing analysis. The remainder of this paper will discuss those changes in more detail.

FUNCTIONAL ANALYSIS

14. We have found that the introduction of new equipment or movement into new technology rarely impacts on only one trade or classification, rather it can/cause a whole series of changes and blurr the division between MOCs. It may be necessary to introduce a new MOC, change several and/or delete an existing MOC. In order to identify the changes necessary we are currently examining functional areas in multi-stream surveys. In these surveys we treat a group of MOCs as one large MOC and have all the job incumbents complete exactly the same questionnaire. An example, is an analysis currently underway which treats eleven air technical MOCs as one " airplane fixing " function. Analysis of tasks, knowledge and skill determine if the current structure is sound or indicate where changes are needed. We are currently examining naval trades and communications trades in the same fashion.

15. During a recent Aircrew analysis we examined an aircraft mission from planning to post mission de-briefing as a single function. Two enlisted and two officer MOCs were involved. Tasks, knowledge and skills dealing with support, ground duties and administration were set aside while we looked at what needed to be done and who could best do it in terms of the mission requirements. The relationships of crew members were sorted out by examining tasks performed, time spent, identification of knowledge needed and level of knowledge, identification of skill used and level of skill and the task involvement (i.e. assist, do, or supervise). We were able to confirm that the crew structure was sound but that in the case of anti-submarine warfare aircraft, pilots were indicating supervisory involvement in many tasks for which they had received little training. Having confirmed that supervisory involvement was in fact required, pilot training was adjusted.

KNOWLEDGE AND SKILL

16. We have incorporated knowledge and skill lists in our questionnaire in much the same manner as tasks. The lists are compiled from training documents and other sources and validated by subject matter experts. We then ask job incumbents to identify the knowledge and skills needed on the job and to rate the level of skill and knowledge required, using a seven point scale. Great care is taken to ensure that the job incumbent rates knowledge at the level the job requires (not that achieved through training or experience). In the case of skills the job incumbent is asked to rate those skills he uses at the level he uses them in his current job.

17. The data collected from job incumbents has been reviewed by subject matter experts and found to be acceptably valid with ratings consistently about half-point over-inflated. It is intended that in at least some surveys to have immediate supervisors provide an assessment of the knowledge and skill requirements of job incumbents in a further effort towards validation.

18. Much work still needs to be done in the analysis of knowledge and skill data. However, thus far we have been successful in substituting knowledge or skills for tasks in the overlap process. Clustering of knowledge or skills produces very nearly the same groupings of jobs as a time diagram giving an added dimension to analysis. Knowledge and skills have also been included in the history file as added variables permitting a group of interest selected from a time diagram to be amplified for more comprehensive analysis by the addition of the knowledge and skills as well as the level of knowledge and skill for that group to the task information.

EQUIPMENT LISTS

19. Equipment lists have been incorporated into our analysis questionnaire for the two-fold purpose of keeping the task list down to a manageable size and to make it somewhat easier to maintain one level of specificity in the task lists. As with knowledge and skills, the lists are gathered from training and technical documents and validated by subject matter experts. A seven point involvement scale is used and benchmarks developed to fit the needs of a particular survey. When surveying equipment operators it has not been necessary to establish involvement in great detail. A simple three point scale of operate, maintain or repair has been found to be adequate and in some cases a yes/no reply indicating some type of involvement has sufficed. For technical (maintenance) MOCs a more specific identification of involvement such as install/remove equipment, perform planned maintenance test, and replace components is needed. As with knowledge and skills equipment lists can be clustered to establish job involvement with equipment or used as added variables to provide added clues to the analysis of a group selected from a time diagram.

JOB DESCRIPTIONS FOR NEW JOBS

20. With the re-equipping of the armed forces and the appearance of new work areas (in some cases whole new jobs), there is a requirement to identify the new tasks that will need to be performed in advance of introduction of the equipment. Once the job is identified it is necessary to decide whether the new job requires additional tasking for current MOC, should be shared by several MOCs or requires the introduction of a new MOC.

21. A simple method of developing a job description for a new job was developed. The methodology is not completely accurate, however it serves the purpose of sorting out trade structure problems and allows training to commence before the equipment arrives. The first step is one of assembling information already available and extracting those tasks which obviously will need to be performed to operate or maintain the new equipment. In the case of a new work station in our new maritime patrol aircraft where the equipment is used to proc acoustic data, it was found that a similar fitment existed in the USN and that an occupational survey embracing the equipment had recently been completed. Th allowed us to extract tasks from the USN survey as a start point to our new tas list. We then added tasks gathered from our own surveys of two MOCs who had so involvement in acoustics processing in our current aircraft. Added to the job description were tasks taken from the aircraft manufacturers literature describ operation of the new equipment and lastly a study completed by NPRDC. The task list produced by our analysis staff was validated by the joint effort of those agencies possessing information on the acoustic system. These agencies includec

new equipment project offices, the acoustic analysis center, equipment standards offices, trials unit etc. - a total of 8 different organizations. The task list (job description) produced has already been used for training and been shown to be accurate.

SUMMARY

22. Our organization is deeply involved in the resolution of several problem areas which include a re-examination of many occupational fields, helping to halt the proliferation of training, and finding ways to introduce new equipment and sophisticated technology in the most efficient and cheapest way possible. Because of a need for quick solutions we have taken a pragmatic results oriented approach. To determine a viable structure we are examining complete career fields, trade families or missions in a single broad questionnaire. We have developed a method of using an Occupational Analysis Survey to gather knowledge and skill requirements and found ways to achieve more comprehensive analysis by the addition of knowledge, skill and equipment involvement data. Finally, we have developed a simple process to produce job descriptions for equipments considerably in advance of its introduction.

THE USE OF CODAP IN NON-JOB ANALYTIC APPLICATIONS

Kennith C. Hogue

Occupational Research Program
Department of Industrial Engineering
Texas A&M University
College Station, Texas 77843

INTRODUCTION

This presentation is concerned with the use of the CODAP programs in analyzing data relative to the Marketing and Distributive Education Co-op training program available in over /00 Texas public schools. Before I get into the data collection and analysis part of the presentation there are two definitions I need to give you.

First, Marketing and Distributive Education, referred to as the DE program, is a program of instruction in sales and marketing designed to prepare high school students for careers in a wide range of sales and marketing jobs, including the ownership and management of businesses engaged in the marketing or distribution of goods and services. It is a Co-op training program in that students are placed at a business location or training station to receive on-the-job training while being reinforced with classroom training in his DE class at school.

Second, placement is defined as a marketing job or training station where a student is employed to receive on-the-job training.

Now, to continue, the Texas Education Agency staff recognized that the continuous growth of the Texas economy and ever changing methods used to distribute goods and services from manufacturers or producers to consumers has caused a realignment of career opportunities in marketing and distributive jobs. This realignment has resulted in new and emerging occupations that need additional placement and training emphasis through the DE programs in Texas. To be sure that DE teachers and school administrators are abreast of these changes and to enable them to provide appropriate curriculum and career counseling services to their students, the Texas Education Agency contracted with the Occupational Research Program at Texas A&M University to conduct a study to identify emerging marketing and distributive occupations.

To identify emerging marketing and distributive occupations, information from businesses and DE teachers was needed. Survey forms were developed for a large sample of businesses and all DE teachers in Texas to accomplish the collection of this information. For the purpose of this paper, only the data from the DE teachers will be presented.

There were two questions asked by the ORP staff and TEA that needed to be answered along with the questions that originally prompted this study. These questions are:

1. Do DE teachers make different sets of placement or have different placement patterns similar to incumbent workers performing different sets of tasks or having different job descriptions?

2. Do DE teachers with similar placement patterns have similar backgrounds.

PROCEDURES

Considering the experience of the Occupational Research Program with the CODAP system it was natural for us to select the data collection and processing methods required by CODAP to answer these questions. We determined that the information we needed was comparable to the background and task sections of a job inventory. We also determined that meaningful performance (placement) and time spent data could be collected.

For the purposes of this study the task listing found in a job inventory was replaced by a listing of job titles with general business classifications replacing duty titles (see Figure 1 on the following page). The background section of the survey instrument is identical to that of a job inventory. The time spent definition was modified to fit this circumstance. Time spent is defined as the time a DE teacher spends placing students, counseling with students, parents, employers and others, preparing and grading student's individual assignments and any other time spent relative to a given job title. A seven point time spent scale was used in this study.

After these modifications were made and appropriate instructions for completing the survey were developed, the survey instrument was printed and mailed to over 600 DE teachers in Texas. Three hundred eighteen or approximately 50% returned the survey. The returned surveys were reviewed for completion and keypunched for entry into the CODAP system. The CODAP programs used were INPSTD, VSETUP, OVLGRP, PRTVAR, JOBDEC, VARSUM, and AVALUE.

PRESENTATION AND INTERPRETATION OF DATA

To answer the first question, "Do DE teachers have different placement patterns?" groups 19, 21, 51 and 110 were selected for this investigation from the clusters created by the OVLGRP program. These groups are made up of teachers with similar placement and similar time spent estimates on placements made.

RELATIVE PERCENT TIME SPENT SCALE 1 Extremely Low; 2 Very Low; 3 Low; 4 Average (Neither Low or High); 5 High; 6 Very High; 7 Extremely High	✓ Current placements 1	✓ Other placements in past 5 yrs 2	Time rate placements 3	✓ Instructional material not adequate 4	✓ Your experience 5	
04.0200 APPAREL AND ACCESSORIES CON'T.						
11 Salesperson, Orthopedic Shoes						10-
12 Salesperson, Shoes						15-
13 Salesperson, Women's Apparel and Accessories						20-
14 Salesperson, Yard Goods						25-
15 Sales Representative, Safety Apparel & Equip.						30-
16 Sales Representative, Uniforms						35-
Add jobs which are not listed and make appropriate checks.						
04.0300 AUTOMOTIVE						
1 Car Rental Dispatcher						40-
2 Inventory Control Clerk						45-
3 Manager of Parts						50-
4 Salesperson, Automobiles						55-
5 Salesperson, Automotive Accessories						60-6
6 Salesperson, Auto Rental or Leasing						65-6
7 Salesperson, Parts						70-7
8 Salesperson, Trailers and Motor Homes						5-
Add jobs which are not listed and make appropriate checks.						
04.0400 FINANCE AND CREDIT						
1 Cashier						10-1
2 Cashier Supervisor						15-1
3 Clerk, Account Information						20-2
4 Clerk, Charge Card						25-2
5 Clerk, Charge Account						30-3
6 Clerk, Commodity Loan						35-3
7 Clerk, Credit						40-44
8 Clerk, Credit Records						45-4
9 Clerk, Disbursement						50-54
10 Clerk, Foreign Exchange						55-59

FIGURE 1

TABLE 1

Number of DE teachers in each group and summary of $\bar{X}$ placements, standard deviations and total placements by teachers in each group.

Group #	19	21	51	110
N	23	33	95	72
$\bar{X}$	16.57	24.8	30.13	57.58
S.D.	3.50	5.40	6.03	14.19
Total Placements	100	152	191	224

The first data summary we reviewed was mean placements and total placements made by each group. As you can see in Table 1, each group docs in fact have different placement patterns at least as far as the number of job titles used for placements are concerned.

Table 2 on the following page presents the general business classification placement patterns for each group. These are comparable to duty level job descriptions in a regular CODAP job analysis. Each general business classification is ordered by the percent of DE teachers making placements in each job title. Also presented is the cumulative time spent on the job titles in each category. As you can see DE teachers do have different placement patterns and they do spend different amounts of time on their placements.

It is interesting to note that Food Distribution ranks first in Group 21, second in Group 19 and is equal to Apparel and Accessories in Groups 51 and 110. Also Apparel and Accessories is ranked first in all groups except Group 21. The first three categories in Group 110 requires 51.90% of the teacher's time. The same categories in Group 51 requires 60.64% of the teacher's time. Although the same three categories appear in Group 19 they have different placement percentages and 65.99% of the teacher's time is spent in these categories. Notice also that in Group 21, teacher's spread their time over more types of placements in that 59.50% of their time is spent on placements in 5 categories.

To answer the second question, "Do DE teachers with similar placement patterns have similar backgrounds," information from the VARSUM program will be presented.

TABLE 2

General business classification placement pattern for Groups 19, 21, 51 and 110 presented in order by the % of DE teachers making placements in each classification with cumulative time spent estimates.

Duty Titles	% of Members Making Placements	Cumulative % Time Spent by All Members
GROUP 19		
Apparel & Accessories	100.00	26.68
Food Distribution	95.65	47.79
General Merchandise	91.30	65.99
Food Service	69.56	72.07
Automotive	69.56	78.30
Petroleum	65.22	82.42
Hdwr., Bldg. Mtrls., etc.	52.17	86.08
Finance & Credit	39.13	88.92
Personal Services	34.78	91.05
Industrial Mrktg.	30.43	93.42
Home Furnishings	26.08	94.85
Advertising	26.08	95.79
Floristry	21.74	97.18
Recreation & Tourism	17.39	98.86
Hotels & Lodging	13.04	99.28
Insurance	8.69	99.56
Transportation	4.34	99.92
GROUP 21		
Food Distribution	100.00	15.62
Apparel & Accessories	96.97	28.52
Automotive	96.97	35.58
General Merchandise	96.97	49.14
Hdwr., Bldg.Mtrls., etc.	96.97	59.50
Petroleum	93.94	66.15
Food Service	90.91	74.72
Home Furnishings	75.76	80.72
Floristry	69.70	84.31
Personal Services	63.64	87.33
Advertising	63.64	90.66
Industrial Mrktg.	39.39	93.98
Finance & Credit	36.36	96.67
Recreation & Tourism	36.36	98.45
Hotels & Lodging	12.12	98.93
Insurance	9.09	99.25
Real Estate	6.06	99.45
Transportation	6.06	99.78
International Trade	3.03	99.92

TABLE 2 CONTINUATION

General business classification placement pattern for Groups 19, 21, 51 and 110 presented in order by the % of DE teachers making placements in each classification with cumulative time spent estimates.

Duty Titles	% of Members Making Placements	Cumulative % Time Spent by All Members
GROUP 51		
Apparel & Accessories	100.00	21.47
Food Distribution	100.00	39.11
General Merchandise	100.00	60.64
Food Service	85.26	67.87
Petroleum	81.05	71.72
Automotive	77.89	75.43
Hdwr., Bldg. Mtrls., etc.	76.84	80.90
Finance & Credit	62.10	84.49
Personal Services	58.94	87.23
Floristry	58.94	89.53
Advertising	54.74	92.06
Home Furnishings	50.52	94.50
Recreation & Tourism	49.47	97.38
Industrial Mrktg.	26.31	98.57
Hotels & Lodging	17.89	99.21
Insurance	6.31	99.41
Transportation	6.31	99.68
Real Estate	5.26	99.84
International Trade	2.10	99.89
GROUP 110		
Apparel and Accessories	100.00	16.35
Food Distribution	100.00	28.81
General Merchandise	100.00	51.90
Hdwr., Bldg. Mtrls., etc.	98.61	59.62
Automotive	95.83	63.76
Food Service	93.05	69.61
Home Furnishings	91.66	74.55
Recreation & Tourism	91.66	80.86
Petroleum	88.89	83.87
Advertising	83.33	86.72
Personal Services	79.16	89.72
Finance & Credit	79.16	93.74
Floristry	73.61	95.74
Industrial Mrktg.	55.55	97.90
Hotel & Lodging	30.55	98.56
Transportation	29.16	99.45
Real Estate	9.72	99.63
Insurance	6.94	99.76
International Trade	5.55	99.84

Table 3 presents how teachers in the sample are distributed within given population ranges. The first thing that should be noted is that Group 21 is a small city (1-50,000) group while teachers in groups 51 and 110 are located in small (1-50,000) and large (301-3,300,000) cities. Teachers in Group 19 seem to be distributed across small, medium and large cities.

TABLE 3

Population of cities where DE teachers in Groups 19, 21 51, and 110 are located presented by % of teachers in population ranges.

Group #	19	21	51	110
1-50,000	47.8	90.6	58.9	47.2
51-150,000	21.7	6.1	10.5	12.5
151-300,000	4.3	0.0	4.2	1.4
301-3,300,000	26.1	3.0	26.3	36.1

Table 4 presents the distribution of teachers across school sizes in the sample. Group 21 represents small schools which is expected since they are located in small cities. Groups 19 and 51 represent middle size schools while Group 110 seems to represent medium to large schools.

TABLE 4

Student population of schools where DE teachers in Groups 19, 21, 51 and 110 are located presented by % of teachers in population ranges.

Group #	19	21	51	110
1-500	13	42.4	10.5	6.9
501-1,500	47.8	42.4	52.6	41.7
1,501-2,500	30.4	9.1	25.3	29.2
2,501-4,500	4.3	3.0	6.3	22.2

Table 5 presents the experience of DE teachers in the DE program. Group 19 seems to be made up of relatively new teachers with 56.5% having 5 years experience or less while the other groups are fairly equal in experience. Incidently, 45% of the teachers in the total sample had 5 years experience or less.

TABLE 5

Summary of De teachers' experience in the DE program presented by % of teachers in 5 year experience intervals.

Group #	19	21	51	110
1-5 yrs.	56.5	39.4	41.0	40.3
6-10 yrs.	34.7	42.5	32.8	34.7
11-15 yrs.	8.8	15.1	20	21.5
16 yrs. up	0.0	3.0	6.1	3.5

Table 6 below presents the experience of DE teachers in public schools including their experience in the DE program. Again, Group 19 seems the most inexperienced group. Group 21 seems to be the most experienced in that 33.4% of the teachers have 11-15 years experience and 33.4% have over 16 years experience. Groups 51 and 110 are fairly close except Group 110 seems to have a litttle less experience when you consider 32.5% have 5 years experience or less.

TABLE 6

Summary of DE teachers experience teaching in public schools presented by % of teachers in 5 year experience intervals.

Group #	19	21	51	110
1-5 yrs.	38.0	15.2	21.2	32.5
6-10 yrs.	21.7	18.2	23.2	21.1
11-15 yrs.	21.7	33.4	28.5	19.4
16 yrs. up	17.7	33.2	27.1	27.0

Task Difficulty as a Function

Of Combat Scenaric

By

Ruth Ann Marco

And

John B. Mocharnuk

McDonnell Douglas Astronautics Company-St. Louis

St. Louis, Missouri

To be presented at the 21st Annual Conference of the Military Testing Association in San Diego, CA. October 15-18, 1979.

INTRODUCTION

The U.S. Army Field Artillery School at Ft. Sill, Oklahoma is charged with the responsibility of training artillery officers in all facets of artillery system performance. Once component of this system is the location of enemy targets and subsequent destruction of these targets through direction of fire by an observer located in a forward position in the combat zone, usually remote from the artillery pieces. The accuracy and rapidity with which the forward observer (FO) is able to perform these tasks have a direct bearing on the outcome of the battlefield situation, i.e., whether enemy targets are destroyed or disabled. Recent advances in battlefield weapons technology and enemy mobility have made the role of the FO even more critical. Serious concern has been expressed regarding the selection of personnel who are best suited to perform tnese tasks and the requisite training necessary to increase the efficiency and effectiveness of the combat artillery unit. McDonnell Douglas Astronautics Company-St. Louis was under contract to the Army Research Institute to identify the critical factors involved in the selection and training of FOs.[1] In the conduct of this research program, a training task analysis was conducted.

DEVELOPMENT OF THE FORWARD OBSERVER TASK ANALYSIS

The primary objective of the FO Task Analysis was the identification of the critical tasks an FO must complete in order to accomplish his mission. In designing the FO Task Analysis activity TRADOC Pamphlet 350-30 Interservice Procedures for Instructional Systems Development; Phase I: Analyze served as a source book and guide. The following section will summarize the procedures used in the FO Task Analysis.

Development of the Initial Task List

An initial task listing was developed by extracting FO and possible FO tasks from pertinent Field Artillery Officer Basic Course (FAOBC) texts and from direct observation of FO training activities.

A second source of information consisted of classroom observations which included the Observed Fire Trainer (OFT) and the BT-33 simulators and field observations of a walking shoot on the East range, a walking shoot on the West range, a shack shoot using the Gun Direction Computer M18 (FADAC), and a mobile shoot. Additional information was derived from reviews of self-instructional audio-visual materials, interviews with counterfire/survey and gunnery instructors, and pertinent Field Artillery and FO literature. Once the tentative lists of FO tasks were developed, and the lists were consolidated, subtasks and enabling tasks were eliminated, and a preliminary task categorization scheme was developed.

[1] The research reported in this paper was conducted for the U.S. Army Research Institute for the Behavioral and Social Sciences under contract DAHC-19-78-0025 which is more fully documented in USARI-RN-79-19, "Selection and Training of Field Artillery Forward Observers: Methodologies for Improving Target Acquisition Skills."

Authentication of the Task List

The preliminary list of 118 FO tasks was reviewed by 14 FAOBC instructors from the Gunnery, Counterfire, and Tactics and Combined Arms departments at the Field Artillery School at Fort Sill. Nine Gunnery Basic Branch instructors, five Counterfire/Survey instructors and five Tactics and Combined Arms Department instructors were interviewed either individually or in groups of two or three. Each instructor was asked to verify the completeness of the FO task inventory, to identify any additional tasks that may have been excluded, and to eliminate any non-FO tasks. Additionally, they were asked to comment on the criticality and difficulty levels of tasks relevant to their instructional areas with respect to both the operational and training environments.

In discussing the impact of task difficulty and criticality on combat and on training, the instructors surfaced a problem related to the interaction of task differences on several rating dimensions with combat scenarios. Several gunnery instructors pointed out that most of the training that was conducted in FAOBC was directed to a general European combat scenario. However, certain tasks such as terrain association and target location can be very difficult in a desert or jungle environment. The type, quality, and recency of the maps can also differ by geographical region or locale. Maps of Africa and the Far East were described as being incomplete, out-of-date, and, in some cases, of too small a scale to be adequately used. Interviews with other FOABC instructors confirmed these combat scenario differences and identified other examples. Thus, a task that is seemingly very easy to perform in a European combat theater may receive very little emphasis or training in FAOBC. When the FO is placed in an operational environment where the task is very difficult to perform, he may experience great difficulty in performing the task if he can perform it at all. Because task performance in the European scenario may not be representative of other possible combat locales, it was decided to examine the effects of combat scenario on task difficulty in the FO Task Analysis.

Validation of the Task Listing

Sixty-nine tasks were retained in the final task list upon completion of the instructor review of the preliminary FO task list. Task validation was then conducted on this task list. In the task validation phase, interviews were conducted with 56 Field Artillery officers who were assigned as FO's or FIST Chiefs attached to operational units, or, who had recently performed in the FO role. Those participating in the FO Task Analysis included: 15 officers from the First Infantry Division (mech) at Fort Riley, Kansas; 15 from the 9th Infantry Division at Fort Lewis, Washington; and 26 from the 2d Armor Division and 1st Cavalry Division at Fort Hood, Texas. The officers were given instruction in how to complete the Task Analysis Form and then were asked to complete it in the presence of the interview team.[1] Additionally, the interviewees were asked to comment on what they

[1] At Ft. Riley and Ft. Hood, interviews were conducted by MDAC-St. Louis personnel or ARI Ft. Sill personnel. At Fort Lewis, the Task Analysis Form completion was supervised by the Division Artillery staff with written instructions provided by MDAC-St. Louis.

felt was the profile of a good FO, what they thought of the new FIST concept, and what their reactions were to their FAOBC training.

Task Analysis Form - The FO Task Analysis structured interview form (Refer to Figure 1 for a sample page of the Task Analysis Form) included the following information:

a. Task - a specific goal directed activity of an FO described by and action verb and an object.

b. Assumed Prerequisite Skills/Training - an indication of whether or not, for this task, prerequisite skills or training were assumed for each task in FAOBC training and the extent that each interviewee possessed those prerequisites.

c. Frequency of Performance during a Combat Exercise - an indication of how often the task is performed in combat; or, in peace-time, during a combat exercise. It was rated on a five point scale that ranged from "never performed" to "performed very often".

d. Time between Job Entry and First Time Performed - an indication of the length of time between completion of training and performance of the tasks on the job. It was rated using the following scale: (1) Task not yet performed; (2) Task first performed more than two years after FAOBC graduation; (3) Task first performed between one and two years after FAOBC graduation; (4) Task first performed between six months and one year; and (5) Task pe-formed during first six months of assignment after FAOBC graduation.

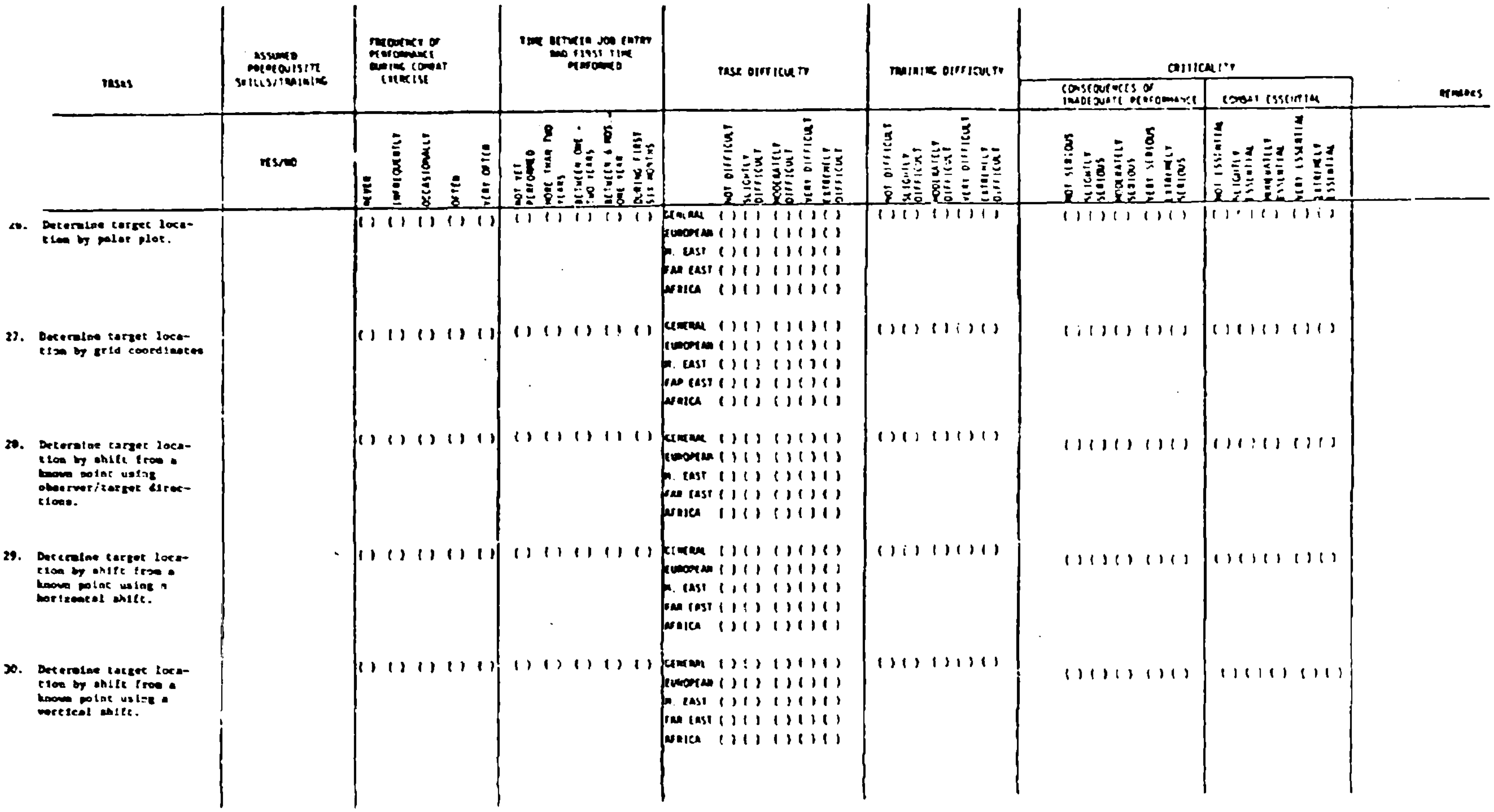
e. Task Difficulty - A measure of the relative difficulty involved in performing the task. It was rated on a five point scale from "not difficult" to "extremely difficult" for five different combat scenarios. The first scenario was a general combat scenario was a general combat scenario which encompassed all possible combat situations. The second scenario included Europe, the third, Far East, the fourth, Middle East, and the fifth, Africa. The scenarios were distinguished along six dimensions: terrain type, ground cover, population density, probable opposition and threat level, air superiority, and map quality. Rating by scenario applied only to task difficulty and, for some tasks, criticality. All other rating categories assumed the general combat scenario.

f. Training Difficulty - a measure of the difficulty involved in learning how to perform the task. It was rated on a five-point scale from "not difficult" to "extremely difficult".

g. Criticality

1.) Consequences of inadequate performance - an indication of the seriousness of probable consequences of inadequate performance. It was rated on a five-point scale from "not serious" to "extremely serious".

2.) Combat essential - a measure of the extent to which the task is essential in combat. It was rated on a five-point scale from "not essential" to "extremely essential".



TASKS	ASSUMED PREREQUISITE SKILLS/TRAINING	FREQUENCY OF PERFORMANCE DURING COMBAT EXERCISE					TIME BETWEEN JOB ENTRY AND FIRST TIME PERFORMED					TASK DIFFICULTY						TRAINING DIFFICULTY					CRITICALITY										REMARKS
																							CONSEQUENCES OF INADEQUATE PERFORMANCE					COMBAT ESSENTIAL					
	YES/NO	NEVER	INFREQUENTLY	OCCASIONALLY	OFTEN	VERY OFTEN	NOT YET PERFORMED	MORE THAN TWO YEARS	BETWEEN ONE - TWO YEARS	BETWEEN 6 MOS. - ONE YEAR	DURING FIRST SIX MONTHS		NOT DIFFICULT	SLIGHTLY DIFFICULT	MODERATELY DIFFICULT	VERY DIFFICULT	EXTREMELY DIFFICULT	NOT DIFFICULT	SLIGHTLY DIFFICULT	MODERATELY DIFFICULT	VERY DIFFICULT	EXTREMELY DIFFICULT	NOT SERIOUS	SLIGHTLY SERIOUS	MODERATELY SERIOUS	VERY SERIOUS	EXTREMELY SERIOUS	NOT ESSENTIAL	SLIGHTLY ESSENTIAL	MODERATELY ESSENTIAL	VERY ESSENTIAL	EXTREMELY ESSENTIAL	
26. Determine target location by polar plot.		( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	GENERAL	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	
												EUROPEAN	( )	( )	( )	( )	( )																
												M. EAST	( )	( )	( )	( )	( )																
												FAR EAST	( )	( )	( )	( )	( )																
												AFRICA	( )	( )	( )	( )	( )																
27. Determine target location by grid coordinates		( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	GENERAL	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	
												EUROPEAN	( )	( )	( )	( )	( )																
												M. EAST	( )	( )	( )	( )	( )																
												FAR EAST	( )	( )	( )	( )	( )																
												AFRICA	( )	( )	( )	( )	( )																
28. Determine target location by shift from a known point using observer/target directions.		( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	GENERAL	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	
												EUROPEAN	( )	( )	( )	( )	( )																
												M. EAST	( )	( )	( )	( )	( )																
												FAR EAST	( )	( )	( )	( )	( )																
												AFRICA	( )	( )	( )	( )	( )																
29. Determine target location by shift from a known point using a horizontal shift.		( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	GENERAL	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	
												EUROPEAN	( )	( )	( )	( )	( )																
												M. EAST	( )	( )	( )	( )	( )																
												FAR EAST	( )	( )	( )	( )	( )																
												AFRICA	( )	( )	( )	( )	( )																
30. Determine target location by shift from a known point using a vertical shift.		( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	GENERAL	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	( )	
												EUROPEAN	( )	( )	( )	( )	( )																
												M. EAST	( )	( )	( )	( )	( )																
												FAR EAST	( )	( )	( )	( )	( )																
												AFRICA	( )	( )	( )	( )	( )																

FIGURE 1. Sample Page from Forward Observer Task Analysis Form

FO Task Analysis Results

FO Task Analysis Summary Data. The responses to all categories of the FO Task Analysis Form were tallied and the percentages for each rating category were calculated.

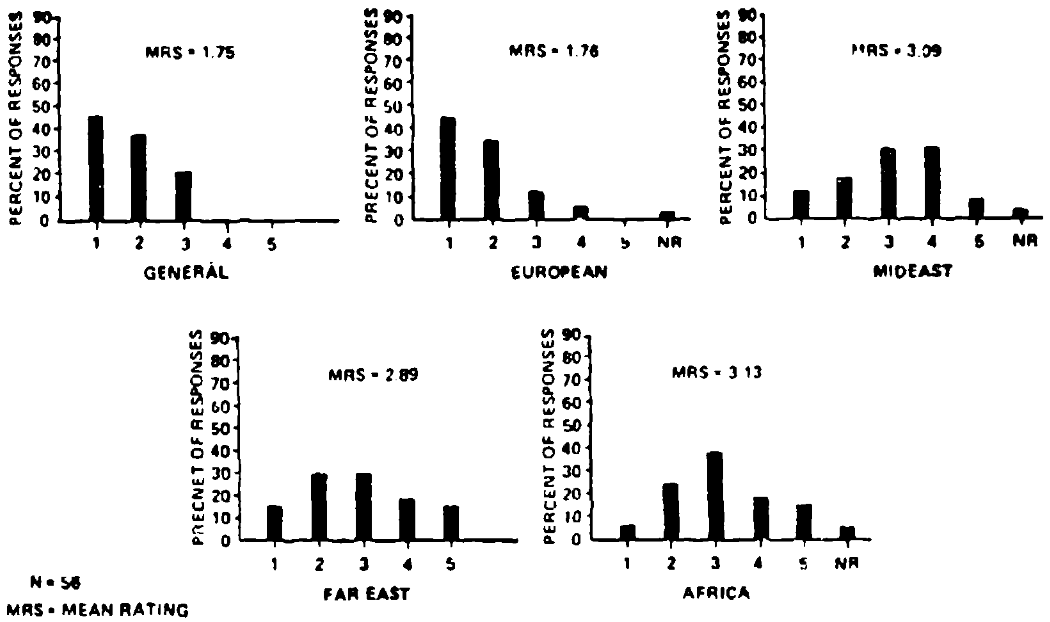
Task Difficulty by Combat Scenario. Similarly, the percentages of response to task difficulty ratings by combat scenario were calculated and summarized. In examining the task difficulty data it became apparent that those tasks which were most affected by terrain differences were the same tasks for which the most variability in task difficulty was found. Of the 69 tasks listed, 25 of them exhibited marked differences in task difficulty across the five combat scenarios. Most of the 25 tasks that exhibited difficulty by scenario differences involved visual/spatial integration abilities which are related to map reading, terrain association, and navigation skills. Furthermore, statistical analysis in the form of chi square demonstrated that these differences were significant.

As an example of the impact of combat scenario on perceived task difficulty the following is a brief discussion of the tasks that demonstrated the most variability in task difficulty by scenario. Each task discussion is accompanied by a graphical representation of the task difficulty ratings for each combat scenario. Task difficulty ratings from 1 (Not Difficult) to 5 (Extremely Difficult) and NR (No Response) are depicted along the abscissa and percentage of response is on the ordinate. Superimposed on each histogram is the mean task difficulty rating score that was calculated for each combat scenario.

a.) Task 4 - Conduct a terrain analysis. As were expected, the task difficulty ratings on this task were highly affected by terrain differences and population density. (Refer to Figure 3-2). These effects were statistically reliable, $\chi^2(20) = 87.17, p<.001$). The African combat scenario, closely followed by the Middle Eastern scenario, was rated as being the most difficult to perform. The African combat scenario encompassed both heavy, jungle-type terrain and large open areas made up of deserts or flatland. The Middle East was represented as being a hilly desert-type terrain with a sparse ground cover. Both the African and the Middle Eastern scenarios were thought to have few population centers and manmade landmarks. Thus, the task of conducting a terrain analysis in either of these two scenarios when there is very little of any substance to aid in the analysis would be much more difficult than terrain analysis in a European combat theater with its varied terrain, dense and numerous population centers, and many manmade landmarks. The Far Eastern scenario, a varied landscape with a moderate number of landmarks and settlements, was rated as being much more difficult than the European scenario, possibly because of the inclusion of heavy jungle amidst the farmland settings. However, the Far Eastern combat scenario is not rated as difficult as the Middle Eastern or the African scenarios. It is interesting to note that the general combat scenario is rated as being almost identical in task difficulty to the European combat scenario. It is the general combat scenario which is taught in FAOBC, and, from our discussions with FAS instructors, the general scenario, in most cases, is the European scenario. The question then arises, are the Field Artillery officers who are assigned to a non-European combat theater adequately prepared to serve as effective FOs?

FIGURE II

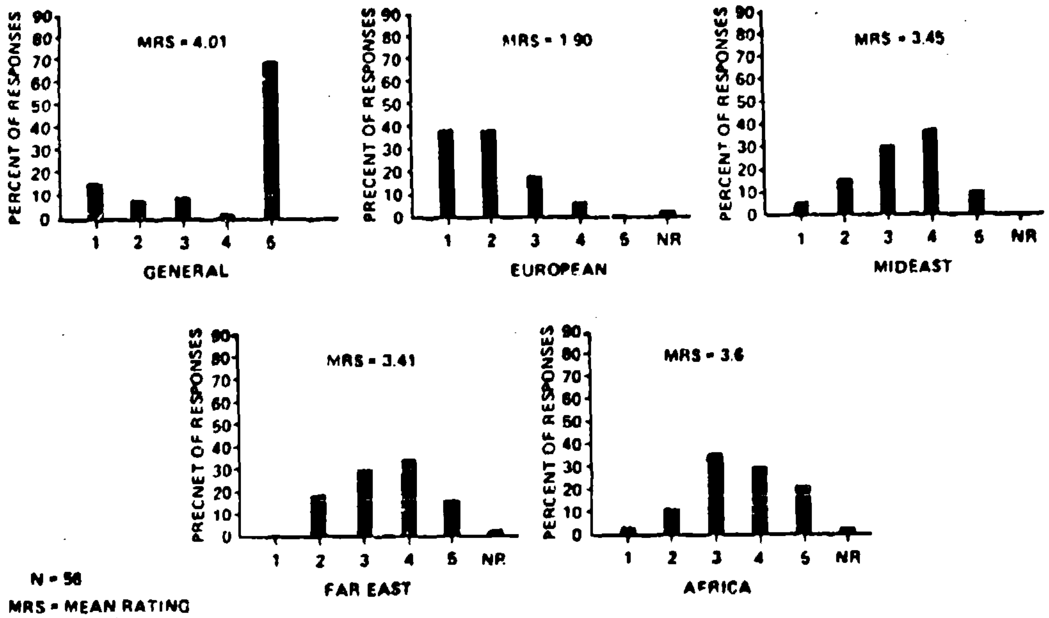
TASK: 41. CONDUCT A TERRAIN ANALYSIS



N = 56
MRS = MEAN RATING SCORE

FIGURE III

TASK: 81. DETERMINE SELF LOCATION BY TERRAIN ASSOCIATION



N = 56
MRS = MEAN RATING SCORE

b.) Task 8 - Determine self-location by terrain association. One of the conclusions of the WSTEA-I study (an earlier investigation of FO performance that was conducted by the Army) was that most experienced FOs have difficulty in self-location skills. Task difficulty ratings on this task tended to confirm the WSTEA-I results. (See Figure III.) These effects were statistically reliable $\chi^2(20) = 168.99$, $p<.001$). This task was one of few tasks in which the task difficulty rating for the general combat scenario was radically different from the European scenario as well as being rated as the most difficult of the five scenarios. For the general scenario, 68% felt that this task was an extremely difficult task to perform. However 74% rated this task as being not difficult to slightly difficult to perform in the European scenario. The African scenario was rated second on task difficulty, followed closely by the Mid Eastern and the Far Eastern scenarios.

When asked to comment on these divergent task difficulty ratings, FAOBC instructors felt the task ratings provided a very good picture of how difficult it is to determine self-location by terrain association. The European scenario would be the easiest because of the large number of manmade structures, landmarks, and roads; and, because of the lack of the same, the African, Mid Eastern and Far Eastern scenarios would be much more difficult. In general, the instructors felt that self-location by terrain association is a very difficult task and many experienced FOs are likely to demonstrate problems on this particular task. The instructors added that the data on the general scenario difficulty rating were indicative of the way it really was and more training time should be devoted to this task because it can be so difficult.

c.) Task 19: Navigate on land by foot.

d.) Task 20: Navigate on land from a vehicle. Figures IV and V depict the difficulty ratings for these two tasks. The effects for both tasks were statistically reliable: for Task 19, $\chi^2(20) = 90.3$, $p<.001$; for Task 20, $\chi^2 = 69.78$, $p<.001$. It is interesting to note that navigating from a vehicle was considered to be more difficult to perform than navigating on foot for each scenario. That is, for the general scenario, navigating from a vehicle was more difficult than navigating on foot. When navigating from a vehicle, there is a much greater opportunity for getting turned around or becoming disoriented, and the ratings clearly reflect this difference. The African and Far Eastern combat scenarios were considered to be the most difficult for both tasks. This was not surprising since both have remarkably varied terrain including thick, dense jungles. Rolling, shifting desert and a scarcity of landmarks led to moderate ratings of difficulty for the Mid Eastern combat scenario. The European and general combat scenarios were rated as the easiest of the combat scenarios on the land navigation task. Thus, land navigation is a relatively easy task in an area where there are roads, many landmarks (natural and manmade), and population centers.

SUMMARY

In reviewing these task scenario and difficulty rating differences, it became apparent that the type of terrain and quality and recency of the maps can severely attenuate the ability of the FO in performing his job. Terrain variance and map quality can add a different level of complexity to tasks that on their own may be fairly simple to perform. With more than half of the selected tasks demonstrating differences in task difficulty ratings for each scenario an important question was raised. How equipped to handle these differences is the FO who graduates from FAOBC with training for only the

FIGURE IV

TASK: 19). NAVIGATE ON LAND BY FOOT

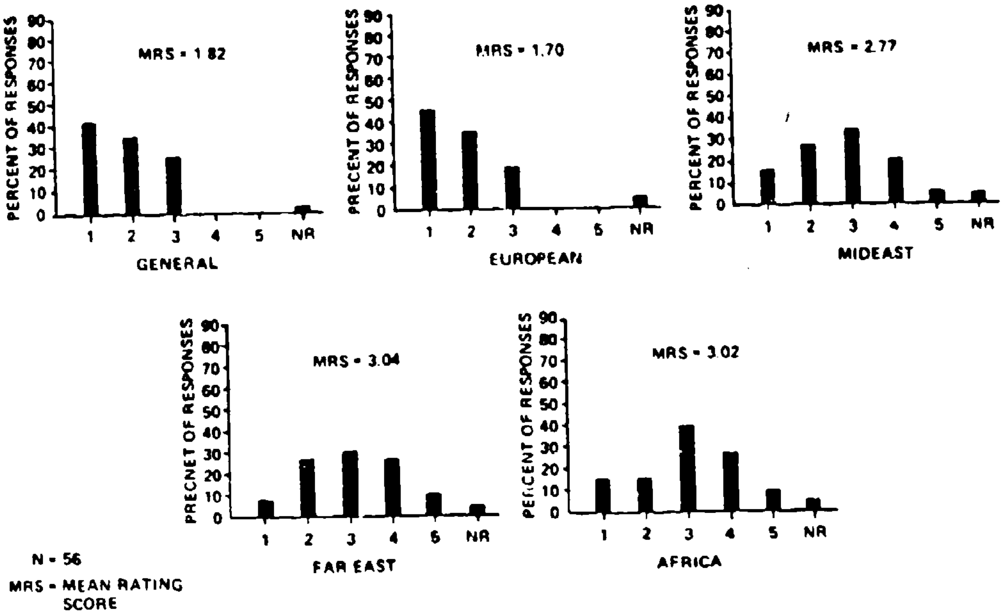
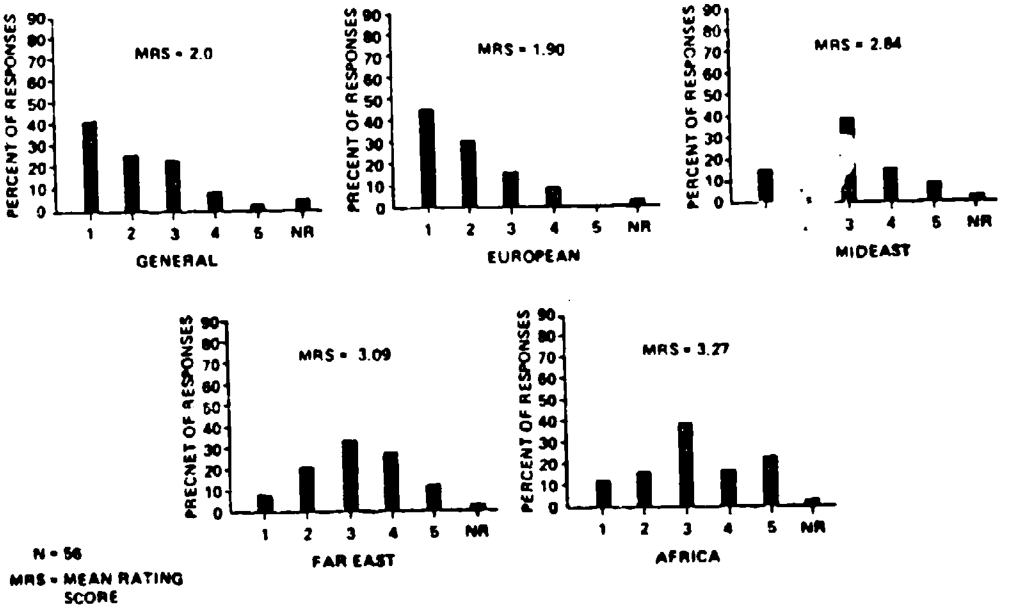


FIGURE V

TASK: 20). NAVIGATE ON LAND FROM A VEHICLE



general combat scenario? The answer might be, he is well equipped to at least deal with the determination of self-location by terrain association since for this task the general scenario was the most difficult. However little emphasis or training time is given to this task because it is assumed that the student in FAOBC has this skill prior to his training at Ft. Sill. Both the FAOBC students and the instructors confirm that this is an incorrect assumption.

Instructors, when queried concerning scenario difference stated that on occasion these differences may be pointed out to the student FO by some of the instructors, some of the time. However, there was no time allocated for the training in development of the techniques in how to perform these tasks for the various combat scenarios. Examination of the COI for FAOBC verified this latter statement.

Two important points must be considered here. One pertains to discrepancies in the required and actual preparation of FAOBC students prior to FAOBC. It is clear that many of the students enter FAOBC with inadequate training in the areas of map reading and terrain association. The second point pertains to the need to consider scenario differences in developing training programs. It is very likely that there will be positive transfer from training for one scenario to application in another scenario; however, the extent of transfer will probably depend on the nature and extent of scenario differences and on the complexity of the specific task. The dramatic differences in task difficulty by scenario which were revealed in the task analysis section make clear the importance of scenario effects. Further discussion of the implications of task difficulty ratings by scenarios is presented in the Training Analysis section.

FAPS[1]

Analysis of the Functions Pertaining to a Job for Better Manpower Structure - Lessons Learned and Ideas of Solution -

Günther Fiebig, Flottillenadmiral, FGN,
Chief, General Armed Forces Office

Rolf-Eckart Rolfs, Fregattenkapitän, FGN,
Ministry of Defense, Armed Forces Staff

Paper presented at 21st Annual Conference of the Military Testing Association 1979

[1] Analysis of the functions pertaining to a job

A. Abstract

In 1971 the "Manpower Structure Commission" recommended a new manpower structure for the armed forces of the Federal Republic of Germany. An instrument had to be developed which registers tasks and analyses functions pertaining to a job. This instrument is called "FAPS".

Its first stage has been in use since 1976. The second stage was started in 1978, but has not been fully developed and programmed yet. Shortages in money, time and manpower require a "concentration of forces". Users want results earlier than planned. They would rather skip the one or the other requirement which do not seem to have priority to their actual problem. This creates extra problems because the instrument has been so designed that each step will be required in order to achieve valid results.

After recalling the functions of the instrument, which has been presented to the MTA in San Antonio in 1977, lessons learned, problems and possible ways out are addressed in order to give a basis for discussion and exchange of experience.

B. 1. Goal of FAPS

The military personnel structure is good when it is "sound", that is to say when

- age structure,
- pay structure,
- grade structure (hierarchy)

are appropriate.

The personnel is satisfied when

- their professional expectations,
- their possibilities to be promoted,
- their activity at the respective working place

correspond to the proficiency and to the ideas of the individual person.

A condition thereof is that every soldier is employed according to his previous education and to his training, on the basis of that education. This therefore necessitates that he is trained according to his capabilities and then employed in correspondence with his knowledge and his skills.

Moreover, possibilities of employment corresponding to his progressing age and to his rising grade must become transparent in the interest of the Armed Forces and in the interest of the man concerned.

In order to reach that transparency, inter alia training series and assignment packages are determined. Since they are job-oriented, but since not every specialty has the same ideal cone structure for age and grade, it is not possible that a soldier - he may be as able as possible - reaches in his training series and assignment package the highest grade of his career. In one training series and assignment package there exists an excess offer of qualified NCOs, who according to their age ought to become sergeants or career soldiers; in another training series and assignment package such a selection basis is missing. The "Funktionsanalyse Personalstruktur" does not solve these problems. There are other systems which compare offer and requirement, propose the transfer to other training series and steer the corresponding training.
But these systems cannot work without fundamental data concerning tasks and requirements pertaining to jobs.

VIEWGRAPH 1

The "Funktionsanalyse Personalstruktur" supplies these data. It registers and analyses tasks and requirements as well as particular strains/working conditions with regard to the jobs. It is not based on a mere description of existing jobs in order to enable a comparison of their evaluation with that of similar jobs, but rather considers individual self-contained tasks. These tasks are standardized in order to enable a comparison with similar tasks pertaining to other jobs, up to those of other services, the medical and health service or the civilian sector. The "Funktionsanalyse Personalstruktur" does not refer the characteristics of requirements immediately to jobs, but first to tasks, combining them then into jobs. It seems that only that approach enables the formation of training series and assignment packages appropriate to the structure, as well as the adding, omission and modification of tasks pertaining to jobs, thus permitting the adaptation of the structure to the offer of personnel and to the military requirements.

The "Funktionsanalyse Personalstruktur" supplies data not only for the formation of training series and assignment packages, but also provides profiles for requirements the application of which facilitates the aptitude test and thus the selection of personnel. The "Funktionsanalyse Personalstruktur" establishes job descriptions which enable an evaluation within the organization.

2. Functioning of the "Funktionsanalyse Personalstruktur"

VIEWGRAPH 2

Requirements which are placed to a man for the fulfilment of tasks are to be measured according to the criteria of

- capabilities
- knowledge
- skills.

The characteristics of requirements will be dealt with later.

"Particular strains/working conditions" - as far as they cccur - are associated with tasks in normal duty (peacetime) and in exercises (state of defense) and classified according to stress.

The function pertaining to a job can be subdivided into several tasks. A task is the purpose-oriented combination of actions. The task is specified, i.e. further subdivided according to subjects and procedures. The specified sub-task is formulated in such a way as to enable a relative sequence according to difficulties and scope (complexity) as well as according to executing responsibility and directing responsibility. Besides, a difference is made between normal duty (peacetime tasks) and exercises (defense tasks).

3. Problems in structuring of tasks

The structuring of tasks unfortunately is not as simple as one might think it to be when looking at the pentagonal illustration (VIEWGRAPH 2). The criteria illustrated there must be enmeshed in such a way that the tasks tailored to each job remain accurately the same.

Experts - in most cases special instructors at schools - first associate functions, subjects/procedures and difficulty/responsibility with sub-tasks in a relative sequence and designate them as "specific forms of tasks", numbered from 1 to 6. These are AUTHORIZED functions of a job.

Line officers and NCOs go to the troops with inquiry papers in order to find out the ACTUALLY performed tasks. By that inquiry on ACTUALLY performed tasks, the AUTHORIZED tasks are examined as to correctness. The inquiry on the ACTUALLY performed tasks does not permit to discern the structure of the tasks concerned. It is intended thereby to preclude manipulations which, for example, might aim at a higher job evaluation. The questions concern the functions which the soldier performs at his job; in this way it shall be controlled whether the pertinent task is performed at all and whether it is done in this form.

In addition, the interrogees indicate how often, under which particular strains/working conditions and with what weighting (importance) these tasks are performed by them.

Apart from the fact that not every job incumbent tells the truth but rather tends to exaggerations in his indications and weightings, the deficiencies in structuring have the following effects:

One of the troopers replied that in the task of

achieving the readiness to move of a tank

he sees one of the functions applying to him:

- testing of aiming device,
- putting into operation of telecommunication installation,
- ammunition replenishment,
- starting of motor.

When he is a driver, he feels that the function of "starting the motor" is assigned to him. The tank gunner marks off the function of "testing the aiming device". The loader and the commanding officer act correspondingly.

The structuring previously programmed on the basis of the AUTHORIZED assignment of tasks now brings about a misrepresentation of the computerized job descriptions. The task of "achieving the readiness to move of a tank" might be assigned to each of these four different job incumbents. The function of "starting the motor" then suddenly appears in the detailed job descriptions of all of them. The job description would become false and nearly turned into ridicule.

This example has been simplified in order to illustrate the risks inherent in an incorrect structuring. There are many more subtle structuring errors which are neither noticed by the experts nor by the interrogators. But the interrogee can discern them. He knows which of the functions of the list he does not perform, and he can mark them off.

The actions excluded and marked in such a way allow optimizing the task structure. Structural changes of tasks or corrections of job descriptions will be possible. But structural changes require expensive follow-on inquiries and for corrections considerable manual and ADP costs and efforts will be necessary.

Another solution would be to determine each individual action instead of tasks arranged in groups and to remain at that lower level. However, the job descriptions would then be very extensive and thus hardly readable. The set of problems could be even greater if it were intended to break the actions down to the actual performance of work. FAPS would then be placed at REFA (basic time and motion study) level. On principle, this would be desirable because of its high degree of accuracy, but it would mean going too far as to extent and objective. Thus, generic terms have to be found in order to get useful and readable results.

A further problem with regard to the task structure is the relative weighting, the sequence of specific forms of a task. The sequence of the forms of a task ranging from 1 to 6 results from their degree of complexity and responsibility assigned to the task concerned. In order to remain clear the various characteristics of complexity and responsibility shall not be discussed in detail. Complexity includes e.g. the degree of difficulty of the procedure or the number and type of items of equipment on which the action has to be performed. Responsibility is subdivided in executing and directing responsibility, the delimitation of which will raise additional problems.

In the following, simplified example the set of problems arising with regard to the sequence, i.e. the valence of the specific forms of a task will be dealt with:

VIEWGRAPH 3

The valence of the specific task forms rises

- 2 is more than 1 because of growing complexity
- 3 is more than 2 because of growing complexity and responsibility
- But will 4 be actually more than 3 merely because of the growing responsibility?

  The degree of complexity is less by far. But which criterion is really more important?
- Will 5 be more than 4, because of its greater complexity and lesser degree of responsibility?

  This play could be continued. It proves the problematic nature of such a judgement as to value.

  However, for mathematical procedures used to evaluate such tasks definite and measurable criteria are required.

  Can this problem be solved in an unbiassed way? Will approximate values not be too subjective, inaccurate and thus incorrect? Since no satisfactory solution has been found so far, the valences of the specific forms of a task which are relative anyway may not be used for the present. This prevents an evaluation by means of ADP. But a compilation of job descriptions will be possible. Job descriptions will be useful if the aforementioned exclusions of action are taken into account. They will be readable if the correct generic terms are used.

4. Problems arising with regard to the analysis of requirements

In Part B (Analysis of Requirements and Evaluation) of the FAPS instrument the requirement characteristics will be the subject of an inquiry, analysis and assignment.

VIEWGRAPH 2

They will be measured by the aforementioned criteria

- capabilities
- knowledge
- skills.

Capabilities are belonging to an individual from birth. They are innate and will remain comparatively stable for life. With their aid and by means of training the individual will gain knowledge and develop skills.

Knowledge, skills and capabilities, pertinent fundamentals (such as regulations, laws etc.) and procedures or methods required for the performance of the task must be assigned to tasks. The items of equipment must be taken into consideration in a more detailed way than required for the specification of tasks.

The requirement characteristics must be clearly defined so that their contents and weighting (level) will remain constant even when assigned to different tasks. Thus, "engine", for example, could not be assigned to the functions of "flying" or "maintenance" with the same meaning, since the knowledge a pilot and a mechanic must have of an engine differs considerably. The pilot must know its function, whereas the mechanic must know its construction from the engineering aspect. Problems arise, if the previously established specific forms of a task are not considered suited any longer and have to be changed. Should the structure be changed? Should use be made of exclusions again (this time excluding requirement characteristics instead of actions)? Or should the requirements already be taken into account when drafting the structure? At this stage the instrument is getting complex to such a degree that the matter would get completely out of hand unless generic terms are formed. These generic terms should be set up in a hierarchical manner e.g.

- tracked vehicle
- tank
- engine.

Which type of system development would be most appropriate for an equally effective performance of inquiry and usage?

There are still many problems to be solved. Difficulties also arise in connection with the definition of knowledge, skills and capabilities. Thus, e.g. field officers have to inquire soldiers in order to find the respective capabilities, and they have to define the capabilities correctly although they are military experts and no psychologists. The profiles for requirements must be compatible with the profiles for qualification, which are needed for selecting candidates at a later date. The qualification will be determined by means of a list of 20 capabilities. Will they be sufficient? Will the classification be appropriate?
Only extensive experience or tests will provide the information required.

C. Prospect

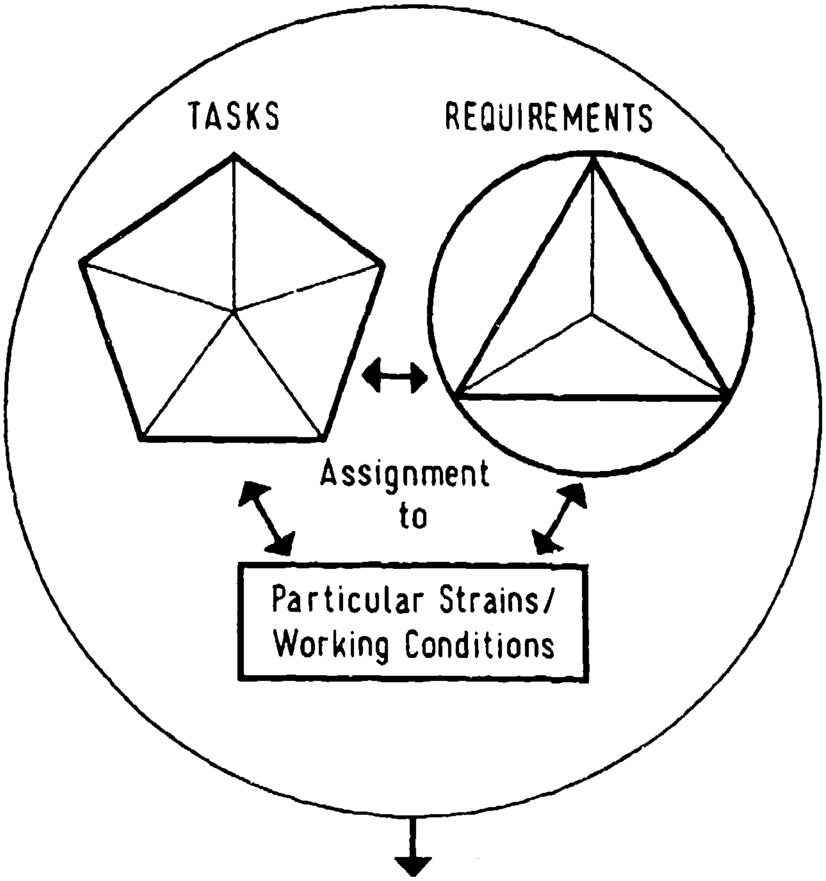
FAPS is a project scheduled to be realized within 6 years. It has been developed and used gradually, step by step. Thus, a gradual utilization of the results seemed to be possible. However, due to the above-mentioned shortcomings no use of these results could be made.

Today methods of correction have to be found in order to be able to make use of the large number of data collected. In addition, trials have to be run by means of which a complete specialty can be checked in order to determine whether this instrument works and where further improvements will be necessary. Only then, the present work can be continued on a large scale.

It is to be hoped that the financial backers are not going to lose patience in this project of many years duration. It would be a pity if this first attempt to collect data for the improvement of the personnel structure had to be abandoned due to lack of money, manpower and time.

VIEWGRAPH 1

FAPS OBJECTIVE



TASKS

REQUIREMENTS

Assignment to

Particular Strains/ Working Conditions

Compilation of

FAPS provides data for:

Training Series and Assignment Packages

↓

Personnel Structure

Profiles for Requirements

↓

Aptitude/ Qualification Tests

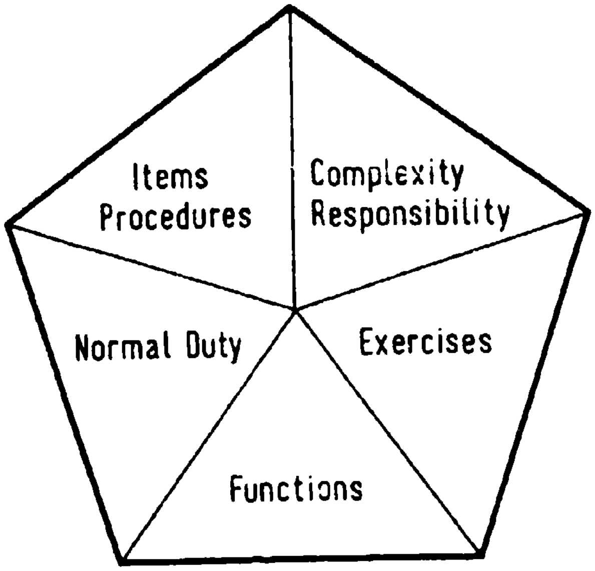
Job Descriptions

↓

Evaluation

VIEWGRAPH 2

TASKS



Items Procedures

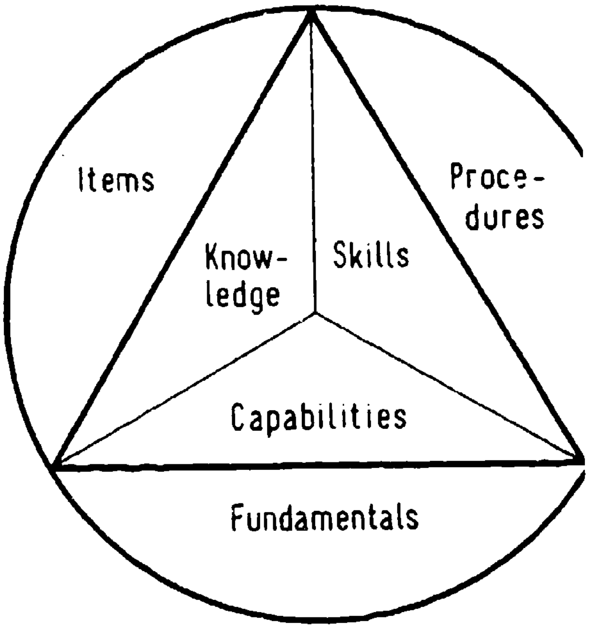
Complexity Responsibility

Normal Duty

Exercises

Functions

REQUIREMENTS



Items

Procedures

Knowledge

Skills

Capabilities

Fundamentals

VIEWGRAPH 3

Specific Form of a Task

1 – modest complexity / little responsibility

2 – average " / little "

3 – great " / average "

4 – modest " / great "

5 -- great " / average "

6 – modest " / utmost "

OVERVIEW OF THE TRADOC TRAINING EFFECTIVENESS ANALYSIS (TEA) SYSTEM

COL Ronald J. Rabin and Dr. Elizabeth Ralls

US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico 88002

The TRADOC Training Effectiveness Analysis System is a management tool for developing and assessing the cost-effectiveness of training subsystems for hardware-oriented total systems and providing training related inputs to the total system acquisition process. The TEA system is applied to the entire life cycle of hardware-oriented total systems such as weaponry, command and control, surveillance and target acquisition systems. The battlefield effectiveness of these hardware oriented total systems is largely determined by how well soldier capabilities match up to the tasks demanded by the hardware. The training subsystem is a major determiner of the quality of this soldier-hardware interface. For this reason the TRADOC TEA System focuses on the training subsystem and its impact on total system effectiveness.

Goals

The TEA system has three goals.

TEA GOAL NUMBER ONE

FIELD COST EFFECTIVE TRAINING SUBSYSTEMS CONCURRENT WITH NEW HARDWARE SUBSYSTEMS

Since 1976, TRADOC policy has stated that training subsystems are to be developed parallel to and in coordination with hardware subsystems. For a variety of reasons, this policy has not been followed. A major part of the TEA work is designed to synchronize and coordinate training and hardware subsystem developments. This is a difficult goal but absolutely necessary to mission accomplishment.

TEA GOAL NUMBER TWO

IMPROVE THE COST EFFECTIVENESS OF TRAINING SUBSYSTEMS ATTENDANT TO HARDWARE SUBSYSTEMS ALREADY IN THE FIELD

Goal number two covers two basic aims. The first is to offset deficiencies which might exist within current training subsystems. The second aim is to reduce the cost of training subsystems that are not deficient as such but are too costly. The interest here is to reduce training cost without lowering Army standards or soldier proficiency, and, if possible, improve on soldier proficiency.

TEA GOAL NUMBER THREE

PROVIDE TRAINING DATA INPUT TO ENHANCE ACQUISITION OF COST EFFECTIVE TOTAL HARDWARE ORIENTED SYSTEMS AND/OR TRAINING DEVICES

The total hardware oriented system has four subsystem components - hardware, training, logistical support and personnel support. The hardware subsystem tends to get high visibility because of its budgetary impact. However, the hardware is only one part of the total picture. The TEA system adds training subsystem data to the picture. Since training data is of fundamental importance to decision makers, TEAs add an important piece of information to the picture of total system effectiveness.

Objectives

The TEA system has specific objectives directed at developing, applying, and refining study methodology so that the system goals can be attained. The objectives of the system are to:

- Help optimize the soldier-hardware and soldier-training subsystem interfaces to enhance battlefield effectiveness.

- Increase the effectiveness of the training subsystem developments process.

- Increase Combat Developments (CD)/ Training Developments (TD) interface early in and throughout the acquisition process.

- Insure that the analysis, design, and development phases of Instructional Systems Development (ISD)[1] are accomplished in a timely manner - before system fielding.

- Improve resolution of Cost and Operational Effectiveness Analysis (COEA) through inclusion of more precise/useful Cost and Training Effectiveness Analysis (CTEA) input.

- Provide baseline data on generically similar systems for inclusion in considerations of developing systems.

- Develop a useful TEA data base.

- Provide for the organization and coordination of the TEA efforts of TRADOC schools/ agencies.

- Minimize duplication of effort and/or redundancy of resource expenditure.

- Provide, within resource limits, readily accessible analytical assistance to TRADOC schools/agencies engaged in TEA work.

These goals and objectives constitute the foundation of the TRADOC TEA System. The next section describes to you the working components of the TEA System and how they fit together.

Description of the System

The Five different types of TEAs that form the core of the TEA System are defined as follows:

- *Cost and Training Effectiveness Analysis (CTEA)*: A continuous, systematic evaluation process conducted during the acquisition cycle of a hardware oriented system to develop training subsystems and training inputs to the Cost and Operational Effectiveness Analysis

[1] See TRADOC Pamphlet 350-30, Interservice Procedures for Instructional Systems-Development, 1 August 1975.

- *Initial Screening Training Effectiveness Analysis (ISTEA):* A systematic study conducted on a fielded hardware oriented system to determine if there is a significant gap between the design effectiveness (E_D) and the actual effectiveness (E_A) of the hardware oriented system.

- *Training Subsystem Effectiveness Analysis (TSEA):* A systematic study conducted to determine if the existence of a significant performance gap is partly or entirely due to the training subsystem.

- *Training Developments Study (TDS):* A systematic study conducted to develop a fix for a training subsystem found to be deficient and/or too expensive and to develop training devices.

- *Total System Evaluation (TSE):* A systematic evaluation of the hardware, logistics support, and personnel support subsystems of a fielded hardware oriented system that is conducted when it has been determined that the training subsystem is neither the sole nor primary cause of a significant performance gap.

There are two ways of classifying the types of TEAs. The major way is to divide them according to the phase of the life cycle with which a given TEA is associated. CTEAs are done on developing hardware-oriented systems and, thus, cover the acquisition phase. ISTEAs, TSEAs and TSEs are done on fielded systems and, thus, cover the post-implementation phase. TDSs can be associated with either the acquisition or the post-implementation phase of the life cycle.

The other division between TEA types has to do with whether or not the relationship(s) of cost and effectiveness are examined. CTEA, TDS, and TSE include study of the cost and effectiveness factors, whereas, ISTEAs and TSEAs typically do not address the question of cost. They do, however, address effectiveness/proficiency.

The basic logic or sequence of events in the TEA system is discussed in terms of the distinction between TEAs for developing systems and TEAs for fielded systems.

TEAs for Developing Systems

Formulation of ways to train users to operate, employ, and maintain a hardware system must begin when the hardware system is first conceptualized. Training implications should be part of the Mission Element Needs statement (MENs). Frequently, a number of ways of training can be identified, and these alternative methods will differ in cost and/or effectiveness. The Cost and Training Effectiveness Analysis (CTEA) evaluates the cost and effectiveness of alternative training approaches as they are being formulated to support the developing total hardware-oriented system. The CTEA is a continuous evaluation process that extends from conceptualization of the hardware system to the point in time when the system is fielded. Thus, the CTEA provides the basis for comparing and refining alternative training methods as the hardware is being developed and provides the basis for a final recommendation on the preferred training subsystem alternative for the system. In addition, CTEA data concerning soldier capability and hardware demands provide direct input to total system COEA. Of importance in this regard is determining through analysis, if hardware driven (H_D) tasks surpass the soldiers' capability (S_C) to perform. If H_D is greater than S_C, the analysis must go deeper to see whether or not selection criteria can be developed and feasibility applied to attenuate the imbalance. Should this fail, either the hardware must be redesigned or grave risks accepted.

All developmental processes for the hardware system must be tied together. That is, the training developments (TD) processes must be accomplished both parallel to and in coordination with combat developments (CD) processes. Evaluations of all the subsystems must be coordinated and reported concurrently. At each decision point (or milestone) of the acquisition process, cost, operational effectiveness, and training effectiveness factors must be considered and the most preferred alternatives retained for continued development. The end result of tying CD and TD processes together will be a cost effective total system, with emphasis placed on effectiveness.

TEAs for Fielded Systems

There are several reasons for continuing to evaluate the total hardware oriented system after it has been fielded. Sometimes when the system is fielded, we become aware of less than perfect decisions made during system development. Also, after fielding, a system often does not retain exactly the same characteristics that it had when the production decision was made. As technology advances, there are changes in management practices and employment tactics. Soldier capabilities and national priorities change too. These changes and new perspectives make it imperative that we continue to evaluate the system, particularly the aspects of soldier-hardware and soldier-training subsystem interface, after the system has been fielded.

To start the post-fielding evaluation process, we must determine if the system is continuing to live up to its design expectations (or design effectiveness) now that it is in the hands of the intended user. Even if feedback from the field appears to give no indication of a performance gap between actual effectiveness (E_A) and design effectiveness (E_D), the proponent should conduct a screening analysis to determine in an unbiased, scientific manner the actual relationship between E_A and E_D. This screening analysis is called the Initial Screening Training Effectiveness Analysis (ISTEA).

Although a first ISTEA may find that E_A equals E_D or comes close, the E_A/E_D relationship must be checked periodically. Why? As mentioned earlier, equipment modifications are almost certain to occur and there may be changes in employment doctrine (an important ingredient of E_A), characteristics of the user soldier population, training subsystems, and the training environment. These changes or others could impact on the actual effectiveness of the total system.

If a performance gap is found and is wide enough and important enough to warrant further analysis, the next step is to isolate the cause of the gap. Training problems are typically the suspected cause when a performance gap is found. A Training Subsystem Effectiveness Analysis (TSEA) is done to scientifically determine if the suspicion about training is founded or not.

If it is determined that training subsystem deficiencies are partly or entirely responsible for the performance gap, a Training Developments Study (TDS) is conducted to develop a fix. If the gap is not partly or totally related to training subsystem deficiencies, the decision must be made to either accept the risk and live with the gap or continue efforts to isolate and find a fix for the problem. If the decision is to identify the problem, a Total System Evaluation (TSE) must be done. The TSE is done to determine what other components of the total system require change, the extent and cost of the change, and whether or not to implement the required change.

Once subsystem changes have been implemented, ISTEAs must be done to assess the impact of these changes on the effectiveness of the total hardware system (i.e., the E_A/E_D relationship). A change in any one of the subsystems will impact the other subsystems and, thus, the system as a whole. If we find that the performance gap has been eliminated or narrowed to an acceptable level, we simply continue monitoring the system.

It is possible, however, that a performance gap may still exist. That is, it is possible that the fixes that were implemented have introduced unanticipated problems. Proper planning and analysis can certainly minimize the probability of unanticipated problems occurring. However, the possibility should be considered and appropriate analyses conducted to identify problems.

Summary

What has been described thus far is the basic logic or flow of events in the TRADOC TEA System. CTEAs, and sometimes TDSs, are associated with developing hardware-oriented systems. The sequence of events for fielded hardware-oriented systems is:

- *First, do an ISTEA to determine if there is a significant performance gap.*
- *If there is no significant performance gap, continue to monitor the system.*
- *If a significant performance gap is found, next do a TSEA to determine if the source of the gap is due partly or entirely to a training subsystem problem.*
- *If training is isolated as the problem, do a TDS to determine the fix that needs to be made.*
- *If the problem is other than a training problem, accept the risk of living with the gap or do a TSE to find what other subsystem is at fault and develop a fix for that subsystem.*
- *If a fix is made, regardless of whether it is a training subsystem fix or a fix of another subsystem, monitor the system by means of an ISTEA.*

Further TEA Relationships

If we look more closely at the reasons for initiating TEAs, we find that some subtle variations in the basic logic or flow of events of the TRADOC TEA System are built in. These subtle variations reflect the flexibility of the TEA system to accommodate the various events that can occur in the life cycle of the total hardware-oriented system. Following is a discussion of these additional characteristics or subtle variations in the TRADOC TEA System. The characteristics are described in terms of particular relationships between the different types of TEAs.

The CTEA-TDS Relationship

The relationship between CTEAs and TDSs points to some interesting characteristics of the TEA System. CTEAs and TDSs are procedurally very similar. Both have the function in the TEA System of developing training alternatives based upon soldier capability and tasks demanded by the hardware. Both require costing of the training alternatives that are developed. Both require that effectiveness measures in terms of soldier proficiency be included in the study.

CTEAs and TDSs are similar in planning logic. The relative merits of training subsystem alternatives for a developing hardware system must be evaluated in terms of variable cost and variable effectiveness. Similarly fixing the training subsystem of a fielded hardware system must be predicated on analyses using a variable cost, variable effectiveness approach.

While both CTEA and TDS must use a variable cost, variable effectiveness planning logic, CTEAs and TDSs differ in two important ways. First, CTEAs deal with the entire training subsystem while TDSs deal only with fixes (i.e., changes or modifications) to a training subsystem. Second, CTEAs are accomplished only for developing systems, while TDSs are associated with either fielded or developing systems. Because a TDS is conducted to develop a fix in a training subsystem,it is the proper analytical tool to be used in developing a training device, whether the training device is for a specific fielded or developing system or is a generic device appropriate for more than one type of hardware system. A TDS may be done to develop a training device for a hardware system that hasn't yet been fielded when development of the training device is initiated late in the acquisition cycle, after completion of OT I. In this case, training device development is not part of the CTEA process itself. (When development of a training device is done as part of the training subsystem development for the hardware system, the training device development is considered to be system related and a part of the CTEA process.)

The CTEA-ISTEA Relationship

ISTEAs are fundamentally designed to answer the question of whether there is a significant gap between the design effectiveness and actual effectiveness of a hardware system. However, the data derived from an ISTEA are also useful to the proper conduct of a CTEA and the proper formulation of a Mission Element Needs Statement (MENS). In a steady state, 10 or 15 years from now, mutually supporting data of this kind will be available to us. For now, when a CTEA is conducted for a newly developing system replacing a generically similar fielded system, an ISTEA must be done on the generically similar fielded system as is discussed in subsequent parts, ISTEA methodology also provides the bases for deriving soldier capability (S_C).

ISTEA, TSEA, and TDS

The analysis of a fielded system need not begin with an ISTEA. If there is strong evidence that a significant performance gap exists, a TSEA may be done as an initial step in analysis. If a TSEA is performed under these conditions, those data elements usually collected and analyzed in the ISTEA must be incorporated into the TSEA data collection plan.

NOTE: Whether an ISTEA or a TSEA is done first, the underlying assumption initially is that there may be a training problem. Training may not be at fault; nevertheless, it provides a starting place for an analysis of the fielded system.

A TDS is also dual purpose. Even though routine follow-up checks on a fielded system may reveal that no performance gap exists, you may still be required for one reason or another to cut training costs. TDSs are designed to fix training subsystems that are too expensive (e.g., develop an alternative to ammo because of its high cost). In such cases, TSEA data must be collected in order to identify the elements of the training subsystem that are to be focused on in the TDS. The TSEA should identify which elements of the training subsystem could be modified or eliminated to cut costs without lowering Army standards or training effectiveness, and, if possible, improve training effectiveness in the process. The TDS is done to develop the fix that maintains or improves training subsystem effectiveness, but does not degrade it.

ISTEA's Relationship to TSEA, TDS, and TSE

The basic logic of the TEA System suggests that an ISTEA is always followed by a TSEA. This is not so. You can follow an ISTEA by a TDS or a TSE, depending on the nature and detail of the ISTEA data. That is, you may follow an ISTEA by a TDS if the ISTEA data pin down the performance gap problem to a particular part of the training subsystem. Similarly, an ISTEA may be followed by a TSE if the ISTEA data are detailed and explicit enough to rule out the possibility that the performance gap is due to a training subsystem problem. However, since ISTEAs will more than likely provide only gross indicators of problem areas, ISTEAs usually require follow-up by TSEAs. Remember, the primary purpose of ISTEAs is to determine the presence or absence of significant performance gaps. Identifying the source of the gap is a secondary purpose.

Summary

The TEA System is a continuous analytic process that begins with conceptualization of the hardware subsystem and ends only when the system is withdrawn from the inventory. Emphasis in the TEA System is on analyzing the training subsystem and improving the cost and effectiveness of both the training subsystem and the total system. There are five types of TEAs. Figure 1 displays how each of the TEAs flow together in the system. The "stop" blocks in the figure mark the points where analysis is terminated because a fix is not possible or is deemed unnecessary.

Basic TEA Concepts

This section explains the concepts that are basic to the TEA system. It begins with a discussion of the factors that make up training effectiveness and explains the TEA system focus on the soldier proficiency factor. In subsequent parts of this section, concepts that are basic to the TEA system are explained in terms of their implications for proficiency.

Effectiveness

The TRADOC TEA system focuses on assessing the impact of training on the effectiveness of hardware-oriented systems. Hardware oriented systems are effective in battle only to the extent that operation of the system accomplishes the functional objectives of that system and meets mission needs. Training subsystem effectiveness is a central concern of the TEA system because training determines in large part how well soldiers will be able to operate and maintain a piece of hardware. While the proficiency of individual soldiers or crews with a particular piece of hardware contributes to but does not solely determine the system's "battlefield effectiveness," it is important to recognize the very important role that individual soldier or crew proficiency does play in the overall effectiveness of the hardware in battle.

The TEA System focuses on soldier proficiency as a measure of training subsystem effectiveness and an important contributor to overall battlefield effectiveness. It also focuses on soldier capability (S_C) to perform and the relationship of that capability to design and actual system effectiveness. These latter elements are of major concern to the COEA process. Soldier proficiency and capability will be discussed later in this section. First, the concept of performance gap must be discussed.

Performance Gap

Performance gap is a key concept in the TRADOC TEA System. A performance gap is the measured difference between a hardware system's design effectiveness (E_D) and its actual effectiveness (E_A). E_D is the expected level of hardware system effectiveness based on an analysis of the parameters of the hardware system. E_D assumes optimally proficiency operators. E_A is the measured effectiveness of the hardware when operated by soldiers in the real world. E_A is based on objective measurement of what soldiers actually do. Soldier proficiency sets the achievable upper bounds for total system effectiveness. In addition to being enlightening about the actual effectiveness (E_A) of the total system, data on what soldiers can do gives

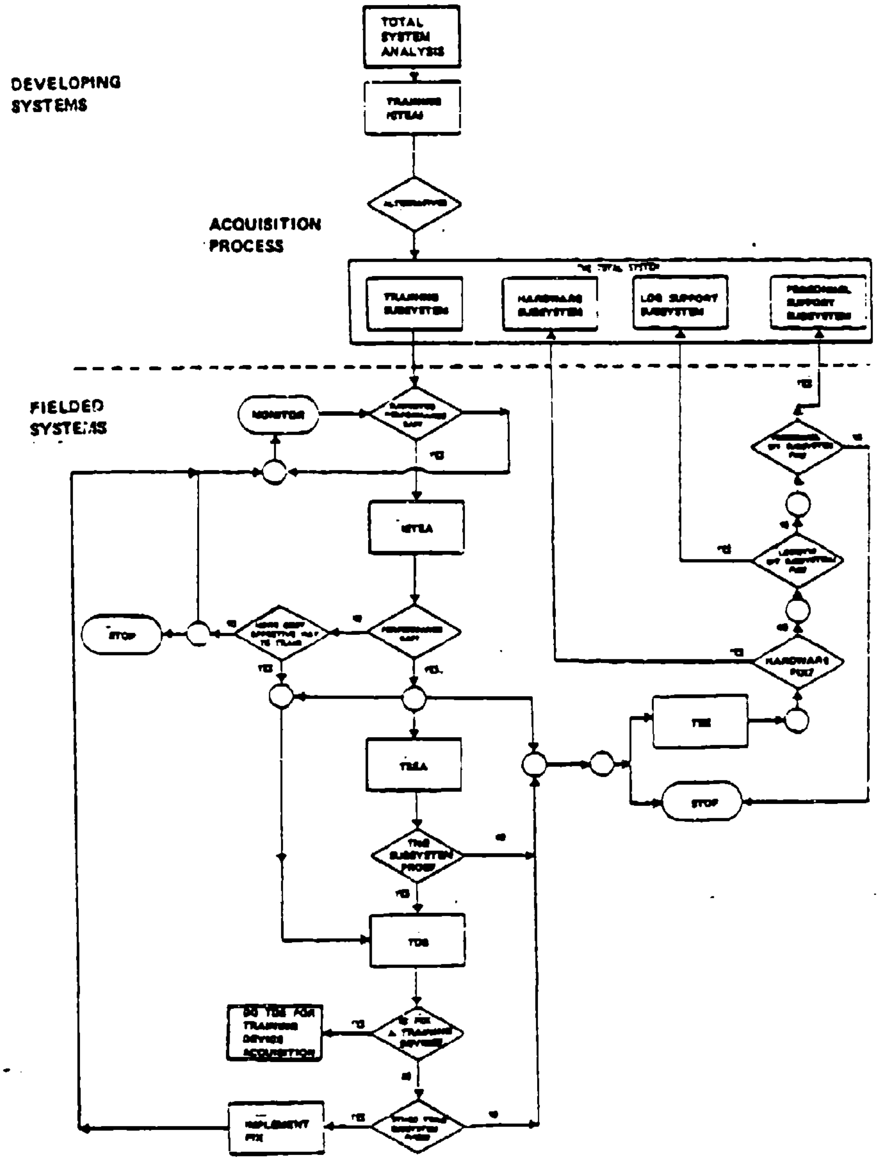


Figure 1
TEA OVERVIEW

insight into not only the effectiveness of the training subsystem, but also their potential with regard to generically similar systems in development.

Sometimes data on the actual performance of soldiers using the hardware is not available, as in the case of the early stages of CTEAs done on developing hardware systems. When this is the case, effectiveness must be conceptualized. Conceptualizing effectiveness involves drawing together knowledge of the design of the system being developed, soldier profiles, the characteristics of generically similar systems in the field, information that can be collected from the contractor and/or trainers of similar systems (e.g., front end analyses and tryouts of the system), etc., and making inferences about what E_A might be like. Thus, conceptualized E_A is inferred from information derived from a number of difference sources.

Soldier Proficiency (Five Factors)

Demonstrated soldier proficiency is the measure of how well a soldier can perform his job. Within the TRADOC TEA system, training subsystem effectiveness is assessed in terms of soldier's performance with the hardware. Soldier performance with the hardware provides a measure of soldier proficiency.

SOLDIER PROFICIENCY IS A PRIMARY
MEASURE OF TRAINING
SUBSYSTEM EFFECTIVENESS

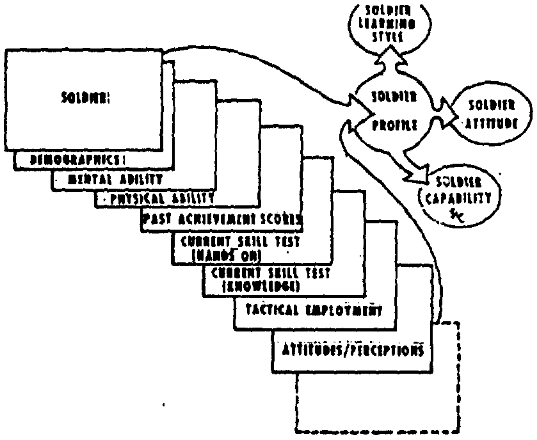
A properly designed, validated, and controlled test will give us the best available estimate of how proficient soldiers/groups of soldiers are. *The problem is not just to identify the existing level of soldier proficiency but to also find out why the proficiency level is what it is.* Identifying the reasons for the proficiency level being what it is, involves determining the factors that underlie proficiency and how they tie together. The factors are many and their relationships are complex. TRADOC has identified the following five proficiency factors as being most intimately related to soldier proficiency.

- SOLDIER
- TRAINER
- TRAINING SUBSYSTEM
- HARDWARE SUBSYSTEM
- TRAINING ENVIRONMENT

The following paragraphs outline the major components of each factor.

Proficiency Factor 1: Soldier

Although all soldiers receive training geared at developing specified levels of proficiency, certain soldier variables sometimes not accounted for play a large part in determining the level of proficiency that is actually attained. One of these is soldier capability. Soldier capability refers to the capability or potential for learning that a soldier brings to the training situation. It is the fundamental variable in the TEA System. Two other variables that play a strong role in determining the proficiency level actually attained are learning style and attitude. Several kinds of data on soldier characteristics must be probed to develop a better understanding of soldier capability, learning style, and attitude. Figure 2 gives some of the various aspects of collectable data that would contribute to developing a soldier profile. The soldier profile provides as complete a picture as possible of three major variables (soldier capability, learning styles, and attitudes) that contribute to actual proficiency.



NOTE: the dotted line is to show the list is not all inclusive—we are on a learning curve.

FIGURE 2 - SOLDIER VARIABLES

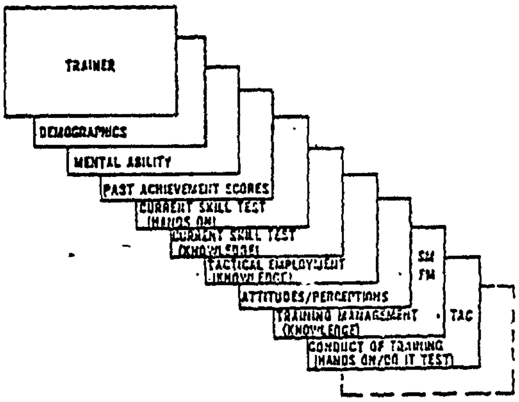

TRAINER
DEMOGRAPHICS
MENTAL ABILITY
PAST ACHIEVEMENT SCORES
CURRENT SKILL TEST (HANDS ON)
CURRENT SKILL TEST (KNOWLEDGE)
TACTICAL EMPLOYMENT (KNOWLEDGE)
ATTITUDES/PERCEPTIONS
TRAINING MANAGEMENT (KNOWLEDGE)
CONDUCT OF TRAINING (HANDS ON/DO IT TEST)
SM
FM
TAC

FIGURE 3 - TRAINER VARIABLES

Proficiency Factor 2: Trainer

The Trainer implements/conducts training. The trainer determines to a significant extent the effectiveness of a training subsystem or its components. No matter how well analyzed, designed and developed it may be, the success of a training subsystem -- measured in terms of SOLDIER PROFICIENCY -- relies heavily on the trainer. The trainer must not only be considered in the design of a training subsystem, but also in the evaluation of the subsystem's effectiveness. The question which must be answered is: "Is the trainer competent both in the subject matter and in the art of instruction?" The initial areas for probing trainer variables that impact on training effectiveness are shown in Figure 3.

Proficiency Factor 3: Training Subsystems. The TRADOC TEA System defines training subsystems as:

> COMPLETE PACKAGES PUT TOGETHER BY TRAINING DEVELOPERS USING PHASES I, II, AND III (ANALYZE, DESIGN, DEVELOP) OF THE ISD MODEL AND CONTAINING ALL MEDIA, MATERIELS, MATERIALS, COMBINATIONS AND SEQUENCES THAT TRAINERS NEED TO IMPLEMENT EFFECTIVE TRAINING (PHASE IV OF ISD).

Two kinds of training subsystems. There are basically two kinds of training subsystems: (1) hardware oriented training subsystems, and (2) career-skill oriented training subsystems. The purpose of hardware oriented training subsystems is to develop training packages to train the individual soldier and/or crews to operate and employ the equipment proficiently in combat. The career-skill oriented training subsystems are concerned with improving the effectiveness of soldiers in broad families of skills so that they become more effective combat leaders, trainers, tacticians, and managers. The TEA system centers on hardware oriented training subsystems, but the principles and techniques herein could also be applied to the career-skill oriented training subsystem analyses.

Training Subsystem Effectiveness. Training subsystem effectiveness can ONLY be measured relative to the demonstrated proficiency of the soldiers who have received the prescribed training. Analysis of training subsystems should include investigation of the areas shown in Figure 4. (More areas for analysis will be added as the TEA is refined.) These investigations are essential to establishing whether or not ISD procedures; CMs, SMs, SQTs, and ARTEPS; TEC lessons; FMs, TMs, and TCs; exported POIs, SPAs, etc.; are synchronized and in complete agreement at least with respect to tasks, conditions, and standards.

Proficiency Factor 4: Hardware

The hardware is the subsystem which the soldier is being trained to operate in combat. For developing systems, one key and critical issue which must be resolved is: "Are the tasks demanded by the hardware (H_D) equal to or less than the soldier's capability (S_C)?" Another key issue, related to analysis of fielded systems, is: "Does the actual effectiveness (E_A) of the hardware equal or almost equal its design effectiveness (E_D)?" The H_D - S_C and E_A - E_D relationships are of major importance in the COEA process as well as the TEA system.

Hardware subsystem variables which should be included in the investigation of soldier proficiency are shown in Figure 5.

Proficiency Factor 5: Training Environment

The soldier, the trainer, the hardware subsystem, and the training subsystem come together for the conduct of training within a training environment. A complete TEA cannot be done without examining the training environment. Several factors that are included in the training environment are training management practices, training priority, personnel availability, and training resource availability. The basic factors comprising the training environment are shown in Figure 6.

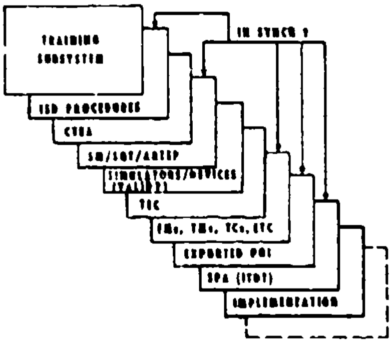


FIGURE 4 - TRAINING SUBSYSTEM VARIABLES

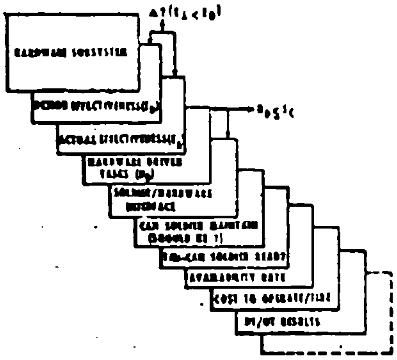

HARDWARE SUBSYSTEM
ΔE(E_A < E_D)
E_D S I C
DESIGN EFFECTIVENESS(E_D)
ACTUAL EFFECTIVENESS(E_A)
HARDWARE DRIVEN TASKS (H_D)
SOLDIER/HARDWARE INTERFACE
CAN SOLDIER MAINTAIN (SHOULD BE ?)
CAN-CAN SOLDIER READY?
AVAILABILITY RATE
COST TO OPERATE/FIRE
DT/OT RESULTS

FIGURE 5 - HARDWARE SUBSYSTEM VARIABLES

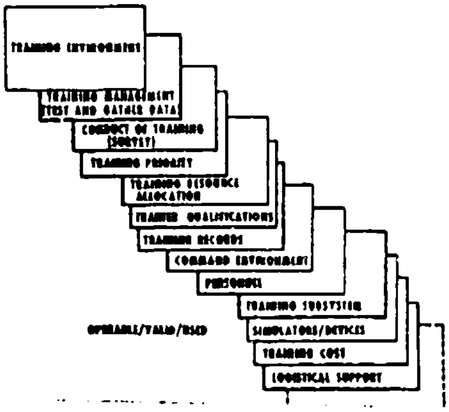


FIGURE 6 - TRAINING ENVIRONMENT VARIABLES

Interaction of Five Proficiency Factors. As shown in Figure 7 it is the interaction of the five proficiency factors that determines the level of proficiency attainable by soldiers.

It is expected that:

> Data collected, analyzed, and interpreted on the Five Proficiency Factors will not only reveal soldier proficiency and allow derivation of soldier profiles but will also provide insights into the factors that underlie existing levels of proficiency. In addition, these Five Proficiency Factors will help explain gaps that exist between E_A and E_D, and provide valuable insights into such important issues as -
>
> - SOLDIER CAPABILITY
> - LEARNING DECAY
> - SELECTION CRITERIA
> - VALIDITY OF SIMULATORS/DEVICES
> - QUANTITATIVE E_A AND S_C DATA FOR COEA

Harmonization of Soldier/Training/Hardware

A key problem that TEA methodology must solve is harmonization of soldier capability, hardware subsystem design, and training subsystem design. This harmonization is needed so that all three elements are optimized. The success or failure of this harmony/optimization has direct, measureable and potentially disasterous implications for the total system COEA.

Soldier - Training subsystem interface. Soldier - training subsystems interface is fundamental to the TRADOC TEA System and critical to proper analysis, design, and development of training subsystems. Soldier - training subsystems interface is a less commonly used term in the Army than is soldier-hardware subsystems interface. Interface of the soldier and training subsystems factors involves developing a soldier profile and then identifying what training approaches or subsystem components are best suited to soldier capabilities, learning styles, and attitudes. The question, "What is the best combination or mix of material, materiel, media, etc., for optimizing soldier training?" cannot be answered effectively by using only the trainers' and/or training developers' "military judgment," "gut feeling," or what have you. Rather, the answer must also be based on as much data concerning soldiers and their characteristics as can be gathered, analyzed, and interpreted. This data must then be combined with knowledge based on military experience to arrive at pragmatic decisions. It is a matter of fine-tuning training subsystems as much as possible so thattheir effectiveness can be maximized in terms of soldier proficiency.

Soldier - hardware subsystem interface. Soldier - hardware subsystem interface involves the match up between hardware demands (H_D) and soldier capability (S_C). The procedure includes doing a task analysis (Phase I, ISD) in order to develop a list of critical tasks that the hardware requires the soldier to perform (H_D) and determining whether the soldier capabilities (S_C) match the hard-are demands. The H_D - S_C relationship must be included in COEAs.

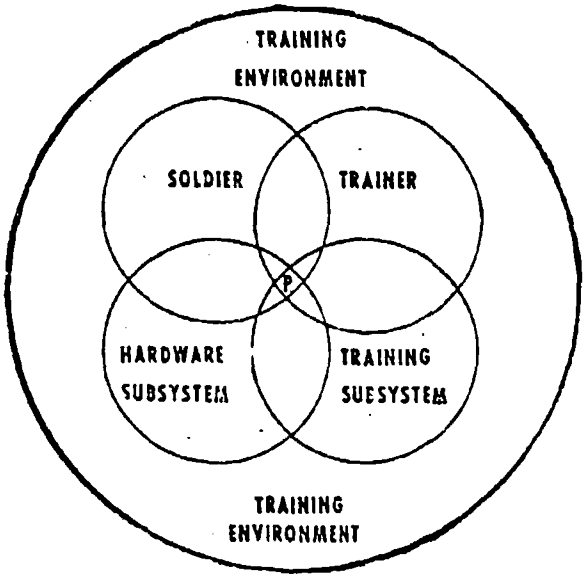
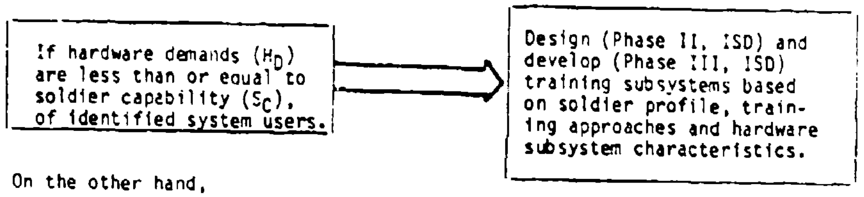


FIGURE 7 - INTERACTION OF THE FIVE PROFICIENCY FACTORS

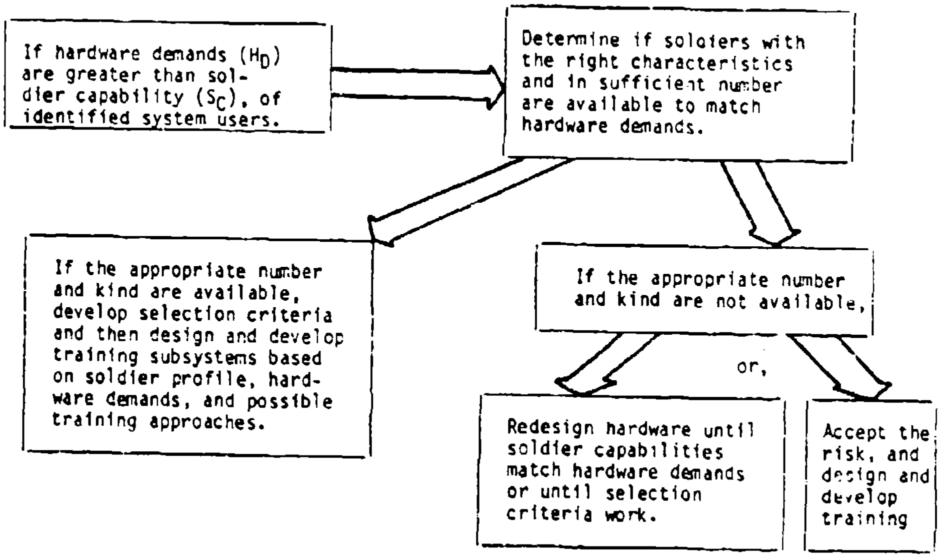
Soldier - training subsystems/soldier - hardware subsystems interface. Development of this three-way interface involves the following logic:



If hardware demands (H_D) are less than or equal to soldier capability (S_C), of identified system users.

Design (Phase II, ISD) and develop (Phase III, ISD) training subsystems based on soldier profile, training approaches and hardware subsystem characteristics.

On the other hand,



If hardware demands (H_D) are greater than soldier capability (S_C), of identified system users.

Determine if soldiers with the right characteristics and in sufficient number are available to match hardware demands.

If the appropriate number and kind are available, develop selection criteria and then design and develop training subsystems based on soldier profile, hardware demands, and possible training approaches.

If the appropriate number and kind are not available,

Redesign hardware until soldier capabilities match hardware demands or until selection criteria work.

or,

Accept the risk, and design and develop training

Cost - Effectiveness Analysis

Of the several possible cost-effectiveness analysis models that could be used, the variable cost, variable effectiveness model is the most appropriate one for comparing alternative training subsystems in TEA. In this model both cost and training effectiveness parameters are allowed to vary. If we properly develop this model considering all relevant system and training cost and effectiveness parameters, it will permit analysts and decision makers to select the overall most cost effective training subsystem. It will also permit development of a more cost-effective total hardware system.

Interdisciplinary Approach

Because analysis of training effectiveness is complex and total system oriented, the aims of the TEA System can only be accomplished by the intercisciplinary approach.

The conduct of TEAs requires the techniques of:

- Operations Research
- Systems Analysis
- Behavioral Science
- Educational Research/Technology
- Training Systems Analysis
- Applied Statistics
- Economic Cost/Analysis
- Military Art and Science

TEA System Imperatives

Changes in procedure are to be expected within the TEA System as it is developed further. The methodologies stated in this paper allow flexibility to accommodate valid changes that are needed to fine tune the system. Any major changes in procedure that would cause radical alternation of the very nature or character of the TRADOC TEA System should only occur gradually. The TRADOC TEA System has some imperatiges that guard against hasty changes in its basic philosophy or nature. The TRADOC TEA System imperatives are that the TEA System must provide for:

- EARLY CONCEPTUALIZATION (BEFORE DT/OT I) OF TRAINING SUBSYSTEMS DESIGN AND DEVELOPMENT
- COSTING OF ALTERNATIVE TRAINING SUBSYSTEMS
- COMPARING EFFECTIVENESS OF ALTERNATIVE TRAINING SUBSYSTEMS
- SELECTING THE BEST TRAINING SUBSYSTEM ALTERNATIVE FOR FULL DEVELOPMENT
- OPTIMIZING HARMONY BETWEEN SOLDIERS, TRAINING SUBSYSTEMS, AND HARDWARE SUBSYSTEMS
- POINTING OUT DISHARMONIES AND RELATED RISKS
- ANALYZING THE E_A/E_D RELATIONSHIP TO DETERMINE IF A SIGNIFICANT PERFORMANCE GAP OCCURS
- ANALYZING THE H_D/S_C RELATIONSHIP TO DETERMINE IF SOLDIERS CAN BE TRAINED TO HARDWARE DEMANDS

Conclusions

The current lack of systemization as reflected in the results produced by past TEAs indicates that a managed, disciplined TEA effort is mandatory. TRADOC Regulation 350-4, 1 June1979, lays the necessary foundations. The genesis of "how to" with regard to the actual conduct of TEAs is found here. As these concepts are expanded, applied in the field, validated, changed, updated, etc., TEAs will become more useful to decision makers and more worth the resource investment. The contents of this paper should convince the reader that properly planned, executed and controlled TEAs are probably the most difficult and complex type of analyses conducted in the Army today. But we must learn how to do them - and do them well. Unless we can get soldiers, hardware, and training together so that we maximize our battlefield capabilities, the first battle may indeed be the last. Technology alone will not win wars or battles. Soldiers trained to use the fruits of technology might. TEAs are the route to better training and more effective/proficient soldiers and, thereby, more effective hardware systems. Cost effective hardware systems cannot be developed until/unless the cost and effectiveness of the total system is scientifically examined. This paper lays the foundation for valuable training inputs which decision makers will find useful because they are objective and reliable.

OUTLINE VULCAN TRAINING EFFECTIVENESS ANALYSIS*

John D. Tubbs

U.S. Army TRADOC Systems Analysis Activity

ABSTRACT

During the past 30 months USATRASANA has been conducting Weapons Systems Training Effectiveness Analyses on the REDEYE Air Defense System, the M60A1 Tank and VULCAN Air Defense System. The purpose is to identify training shortfalls which impact system effectiveness and to improve training such that effectiveness is increased. The VULCAN training effectiveness analysis encompassed 900 gunners from almost every VULCAN unit in the world. Shortfalls, problems, and effectiveness will be described as well as remedies.

*This paper was presented but, due to its unavailability at the time of printing, only the abstract is reproduced here.

JOB/TASK ANALYSIS CHALLENGES AT A NON-MOS-PRODUCING TRAINING INSTITUTION

Robert G. Henderson

Defense Language Institute
Foreign Language Center
Presidio of Monterey, CA 93940

INTRODUCTION

The Defense Language Institute Foreign Language Center (DLIFLC) supports an average enrollment of 2,200 students at any given time and graduates approximately 4,000 students annually from resident programs. At present, twenty-eight foreign languages are taught by an instructional staff of about 450, most of whom are native speakers of the language they teach. (See catalog on following page.)

Our students are predominantly active duty members of the Army, Air Force, Navy and Marine Corps. Adult dependents are permitted to study with their spouses, if they like. A small percentage of our students come from other government agencies, such as the Department of Justice, the State Department and the Central Intelligence Agency. We also have mutual interest contracts with the University of California system and San Jose State University that permits a small number of their students to study at Monterey.

Instructional systems development has been a feature of DLIFLC operations since 1970. In 1976 this concept became much more detailed and structured with the introduction of Interservice Procedures for Instructional Systems Development (IPISD). We have since tried to employ the IPISD concept with only moderate success.

We are challenged, to a large extent, by the heterogeneous nature of the target population, our students. Let me cite some of the variations:

Age: Our students range in age from 18 to 45 years old. Most are under 25.

Gender. We have a growing contingent of female students. Prior to 1972 dependent wives were about our only female students. Nowadays, the ladies represent about 24% of our active duty student population.

The views of the author do not purport to reflect the position of the Army or the Department of Defense.

DEFENSE LANGUAGE INSTITUTE
FOREIGN LANGUAGE CENTER

Current Courses of Instruction

1979

ALBANIAN

ARABIC

BULGARIAN

CHINESE-MANDARIN

CZECH

DANISH

DUTCH

FRENCH

GERMAN

GREEK

INDONESIAN

ITALIAN

JAPANESE

KOREAN

PERSIAN

POLISH

PORTUGUESE

ROMANIAN

RUSSIAN

SPANISH

SWEDISH

THAI

TURKISH

VIETNAMESE

Academic Background. During the years of the draft, nearly all our students claimed some college experience. For a long period the figure hovered at about 14.3 years of formal school prior to entering DLI. That figure has now dropped to about 12 years of prior academic experience. This is somewhat further aggravated by a national decline in reading competency. The first aspect, reduced academic experience, can most likely be attributed to abolition of the draft and introduction of the volunteer army concept. The second observation, declining reading scores, may be due to numerous causes. However, as it affects our training enterprise, we recognize that an individual who does not read well in his or her native tongue is unlikely to read well in a foreign language.

Military Occupational Specialties. We do not train graduates toward a specific military specialty as do most military schools. Our students either possess such a specialty before arriving at Monterey, or acquire one as a result of subsequent training. Our graduates are military personnel who will be required to use a oreign language as part of their primary duty or job. The argest such categories are personnel who will work in the security services, military intelligence, military assistance and advisory groups, attaches and a variety of administrative or liaison jobs. Counting both officer and enlisted personnel for all four services, we have identified over 300 military specialties that use a foreign language to one degree or another in order to perform their jobs. (See display on following page.) As noted earlier, this can involve up to 28 foreign languages or dialects. In general, a dozen or so military specialties account for the majority of our students. Similarly, six languages account for over 80% of our enrollment.

The Language Domain. Not all of our graduates need either the same degree of quality or quantity of foreign language competency. These considerations cut across numerous boundaries that reflect the many ways in which we employ language. We can listen to it, read it, speak it, or write it. And, of course, there is a large non-verbal component. Body language, physical proximity and numerous social conventions are extremely important communicative features in some cultures. Sometimes there is a whole body of occupational jargon that the general native population does not understand, but that same jargon may be crucial to the performance of a military job incumbent. And of course, there are social roles (peers, superiors, subordinates, friends, strangers, etc.), psychological roles (pleasure, anger, confusion, etc.) and registers (colloquial, informal, formal) or modes (persuade, agree, argue, refute, interrogate, etc.). And then there is the corpus of the language itself.

AUTHORIZED LINGUIST POSITIONS

January 1979

ARMY		
Officers	80 MOS'S	467 Positions
Warrant Officers	12 MOS'S	157 Positions
Enlisted	88 MOS'S	2,590 Positions
NAVY		
Officers	31 MOS'S	366 Positions
Enlisted	39 MOS'S	1,054 Positions
AIR FORCE		
Officers	28 MOS'S	296 Positions
Enlisted	59 MOS'S	467 Positions
TOTALS	337 MOS'S	5,397 Positions

The nuts and bolts of a language are its lexicon and grammatical rules. The blueprint for sorting out the nuts and bolts and assembling them coherently is syntax. But any language domain is literally infinite. Not only is the lexical inventory enormous (over 6,000,000 English definitions in Webster's Unabridged Dictionary, for example), but grammatical structure can be similarly forbidding. Though much smaller than the lexicon, esoteric or arcane grammatical features are in a constant state of change lexically, grammatically, syntactically, idiomatically, colloquially, regionally and dialectically. In short, the crucial question facing DLIFLC is: From this literally infinite language domain, in what way can we select most effectively the language content which is relevant to our student population?

ESTABLISHING LANGUAGE PROGRAMS

Instructional development programs at DLI are governed by a Training Development Five-Year Plan (TDFYP), which is updated annually. Any user agency may submit a foreign language training requirement for either resident or non-resident programs. Each service branch and the National Cryptologic Training System (NCTS) has a Service Program Manager (SPM) who acts as liaison between the user agency and the Defense Foreign Language Program (DFLP). The Department of the Army (DOA), acting as Executive Agent (EA) for the Department of Defense (DOD), coordinates proposed programs with the other users and DLI. When programs have been coordinated and approved by DOA, DLI determines the resources and timeframes necessary to develop the program. This is submitted to DOD for review. Programs approved by DOD also authorize DLI to acquire the necessary manpower and fiscal resources to execute the development programs. A major development program, such as construction of a new basic course, may occupy a team of five to seven interdisciplinary in-house experts from three to five years. Training development programs may also be produced by contract with the private sector. With a dozen or more major variables involved, it is difficult to place a dollar figure on the cost of these programs, but it is sizable. On the other hand, the life span of a new basic course of instruction may be fifteen or twenty years with only modest modifications required to reflect changes in the language. The first step in instructional systems development is job/task analysis. In our case, this takes the form of identifying the foreign language linguistic components and the way those components are applied in the conduct of military duties.

ELEMENTS OF THE LANGUAGE DOMAIN

Since our resident courses typically require from six to twelve months of intensive study by our students, the question of which elements to include in the instructional program becomes of acute and pertinent importance. When job/task analysis is applied to vocational skills, the problem may seem difficult, but all the elements of a job or task can be enumerated. Even though the listed hierarchy may be lengthy, it is still finite. Not so with language. And this poses a serious challenge for our institution. Only recently have our job/task analysts really started coming to grips with the problem. The primary objective is to reduce that monumental language domain to manageable proportions. The secondary objective is to include those language functions that are relevant to some or many jobs and ruthlessly exclude the rest from our instructional programs.

As with many enterprises, the theory is simple. It is the execution or application of the concept that is really challenging. To accomplish this, we conduct a long series of collection and analysis procedures using an array of tools and techniques. Our primary targets are current job incumbents using a foreign lnaguage as part of their military duties. Of course, we do not neglect supervisors and staff level personnel, who frequently hold a rather different view on the language and how it is used in pursuit of the organizational mission. In some cases it is helpful to seek out indigenous personnel who work with Americans in the target language. We also like to talk with recent returnee personnel who have held jobs requiring knowledge of a foreign language.

As you might suspect, these groups of people are spread out all over the globe. The first step is to locate them, determine their number, and identify the language involved. There are several ways of going about this. Our best sources of information are periodic reports supplied by the four military branches. Though differing slightly in format, the Army, Navy and Air Force reports are all computer generated, updated about every six months and indicate the command, position title, military grade, numbers of billets and their location. The Navy, uniquely, also indicates the competency levels required for each of the four major language skills (listening, reading, speaking, writing). The Marine Corps report is informal and brief, reflecting their limited language requirements, which are largely limited to interrogation and translation activities.

We have developed a matrix of things we believe we need to know about language utilization in order to develop a new course of instruction that will be related to the world of work and still systematically minimize what the student must learn. The first two dimensions we would like to know something about are criticality and frequency. If it is not essential to be able to say "good morning" in Eskimo it is not likely to be included in the course. On the other hand, it may be critical to express everyday greetings in Basque twenty-three different ways. Or, perhaps only ten of the twenty-three are vital. Those ten forms must be learned by our students. Frequency is also an important element. The more often a part of speech, or a syntactical pattern, is used in connection with military duties, the more important it is for that language element to be learned thoroughly. For example, the manipulation of numbers is inherent in every language. This may be a simple, straightforward operation in Polynesian. But Amharic, for instance, may contain a bewildering array of number declensions that are all equally essential to everyday communication. This relationship must be taken into account in designing a language course.

The ideas of criticality and frequency can be attached to several areas of the language arena. In the past we have made some use of the disciplines of comparative linguistics, psycholinguistics and computational linguistics along with such techniques as contrastive analysis of English with other languages. The practical results have been limited and frequently controversial. More recently we have applied an adaptation of a classification system developed by the Council of Europe. The method is very promising.

To our information matrix, which already includes the dimensions of criticality and frequency, we can add four additional categories. They are topics, settings, social roles and psychological roles. This approach, also applied in a more complex fashion by the British Language Council, is sometimes called the "functional/notional" approach.

Most communicative actions can be associated with a topic, or a series of topics. To view communication otherwise might relegate language to expressions of nonsense in non-literary usage. And language is nearly always purposive. Even a simple event like "good morning" can be viewed as topical. The speaker may be commenting on the weather, indicating that his or her day has gone reasonably well thus far, or may simply be offering a greeting. Which of the three it may be is not always evident without additional information. A topic may be highly specific (Build an H-Bomb in Your Own Backyard), or very general (The Story of Mankind and His Environment in Eight Volumes). Topic is a two-way street. Sometimes we generate the topic and sometimes we simply respond to it. But, in coherent communication, there is always a linkage. In one view it is a simple stimulus response situation. And our responses need not be overt.

Topics occur in settings and settings have several classifications. One, of course, is the environmental locale. Are we at a grand ball, in a business office, around a swimming pool, or in environments different from each other? Another element has to do with physical proximity. Are we eyeball-to-eyeball, across a crowded room from each other, lecturer and audience member, telephone conversants, letter writers and readers, or locked in adjoining prison cells and tapping messages on the wall? Yet another element of the communicative setting is pace, rhythm, or stress. Are we informing, warning, pleading, persuading, arguing, or commanding? Or are we responding to one of these forms of rhetoric? Whether we greet someone, or someone greets us, with "good morning" or "your house is on fire" might make a big difference in responses. Non-verbal communication, or body language, gets involved here too. A thief holding a gun on you and demanding your money accompanied by a warm smile and other friendly gestures is transmitting conflicting messages. I suppose some of us have held "two" conversations while listening to a voice on the telephone and simultaneously making a derisive hand signal to the person at the next desk that non-verbally says a lot about the telephone caller or his message. And we are not likely to be arrested at a baseball game for loudly proclaiming "kill the umpire." Most often, there is a congruity between the setting, the message and the response and, if sender and receiver are within view of each other, vast amounts of information can be exchanged non-verbally.

As we consider some topic in a particular setting we also adopt certain roles, both socially and psychologically. Most of the time, in our own language and culture, we switch roles effortlessly during the course of a single day and often without much conscious thought. When we venture into a foreign language or culture we are out of our natural habitat. The extent to which we can convincingly adapt is often a good gauge of our control of a second language.

Social roles are those that suggest the relationship between message senders and receivers. Are we social peers, superiors, subordinates, or some mixture of these? Are we professionally or technically linked or from widely different parts of the world of work? Are we strangers, kinfolk, intimates, or even romantically involved? Are we near the same age, of very different ages and the same or different sexes? Do we have a doctor/patient, lawyer/client, priest/penitent, or some other relationship? There may be occasions when we do not know what role we should play, or what role is expected of us. But in every case we will use or respond with language and its non-verbal accoutrements very differently for each role.

Psychological roles are markedly fewer, but much more subtle than social roles. Psychological roles operate at the conscious level (what we wish to project) and the unconscious level (what we would rather not project, but may do so without realizing it). Both levels have delicate, but powerful, verbal and non-verbal constituents. Whether overt or not, these roles include the familiar circumstances of joy, anger, anxiety, stress, pride, consternation, interest, ignorance, disagreement, agreement, empathy, hostility, and so on. One important psychological role is humor, or perhaps more accurately, wit. As with other intellectual skills, when one does not possess a certain skill in his or her own language, that individual is unlikely to exhibit that skill in a second language. Thus, if you can appreciate a joke told in a foreign language, you are adapting rather well. If you can tell a joke in a second language, so that native speakers will laugh, you are in beautiful shape.

THE ANALYSIS PROCESS

All right. A quick review. Criticality; frequency; setting; topic; social and psychological roles. That is the matrix to which we would like to tie linguistic elements. The five major tools we use to collect the needed information are questionnaire, interview, observation, analysis and validation. Considerable homework precedes each step.

Part of that homework is the review of job descriptions, position profiles, technical training objectives, training media, job aids and any other printed, recorded or filmed material that will help the job analyst better understand the scope of a particular job or duty. It is not necessary for the analyst to become an expert on all the details of a given job or duty or be knowledgeable in the language involved. But the analyst must be skilled enough to effectively employ the next step in the process, the interview.

The job/duty/position review process is joined with the results of questionnaire responses received from job incumbents to establish the checklists that analysts will use during field interviews. Sometimes that body of information is rather slender. Job and duty descriptions can often be at wide variance with the real world of work. And responses to questionnaires are always fewer in number than one would desire.

Trips to the field are planned on the basis of information extracted from the reports supplied by each service. With surprising accuracy we are able to locate the target population. For a given language, the location of personnel using that language in their everyday duties cuts across service branches and typically are scattered among several sites. Our target audience is not just job incumbents, but their supervisors and key members of staff elements also. A variety of interview techniques are used, all tightly structured so that analyst team members operating at different sites are eliciting responses and gathering data in a uniform manner.

Wherever possible, it is very desirable to observe a job incumbent in the process of carrying out his or her duties. This serves two valuable purposes. First, the interviewer may not be asking all the relevant questions and the job incumbent may not be aware of all facets of the way the foreign language is being used. Second, it provides an early confirmation of details that may not have been considered earlier. As travel patterns crisscross, team members are able to confer frequently and sometimes modify procedures on-the-spot.

The analysis procedure begins even before the traveling team heads back to Monterey. Aside from the interaction with job incumbents, the analysis team has made an intensive effort to collect realia. If a job incumbent must use a Munich telephone book, then we capture a Munich telephone book. If a job incumbent must assist American personnel in filling out some sort of form in the foreign language, we want a copy of that form. If a job incumbent must read a lease or housing rental agreement, we would like a sample copy of that lease or agreement.

Once back home, the team begins the arduous task of bringing all the bits and pieces together. As noted earlier, a primary objective of the analysis process is to identify commonalities among selected military specialties requiring the use of a foreign language. A second objective is to reduce the size of that domain by using the criteria of criticality and frequency. The possibility remains that some important elements have been overlooked, or, things that seemed important to the team, but are not really vital, have been included on our lengthy list of topics, settings and roles. Therefore, we take our final "shopping list" back to the job incumbents.

This is our validation procedure. This activity typically results in a few additions and deletions to our rather lengthy list. It also yields a description of foreign language usage in the world of work in which we can place considerable confidence.

FINAL PREPARATION

The job/task analysts work with a team that includes testing experts, course development experts and subject matter experts. The SMEs are authorities in the foreign language and frequently have extensive teaching experience. It is the job of this team to produce the terminal learning objectives upon which the development of the course of instruction and its associated testing system will be based. We are currently applying these procedures to Arabic, Chinese-Mandarin, German, Korean and Russian. As each project proceeds we are discovering ways to refine the process, but we still have some distance to go in perfecting the job/task analysis operation as it relates to foreign language training.

A FORWARD OBSERVER TASK CLASSIFICATION SCHEME: IMPLICATIONS FOR SELECTION AND TRAINING

John B. Mocharnuk and Ruth A. Marco
McDonnell Douglas Astronautics Company,
St. Louis, Missouri 53166

and

Raymond O. Waldkoetter and John R. Milligan
US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P.O. Box 33066, Fort Sill, Oklahoma 73503

21st Military Testing
Association Conference
15-19 October 1979
San Diego, California 92152

A FORWARD OBSERVER TASK CLASSIFICATION SCHEME: IMPLICATIONS FOR SELECTION AND TRAINING[1]

John B. Mocharnuk and Ruth A. Marco
McDonnell Douglas Astronautics Company,
St Louis, Missouri 63166

and

Raymond O. Waldkoetter and John R. Milligan[2]
US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P.O. Box 33066, Fort Sill, Oklahoma 73503

INTRODUCTION

Based on the results of the forward observer (FO) task analysis in response to a Field Artillery training need, certain probable impacts on training design have resulted and will affect the related training recommendations. As the task analysis developed it was observed that in order to determine the general skill-level requirements or behavioral categories, a task classification scheme could prove useful to interpret task data and findings. Several task classification schemes used in earlier Instructional Ssytems Development (ISD) training efforts were examined for application in the FO task analysis, but they were found to be inadequate because of the restricted scope and range of tasks and jobs that were studied. Most of these task classification schemes (Gagne, 1962) gave little attention, for example, to complex higher-order cognitive tasks which seemed to be a part of the job. After a review of the relevant training literature a new task classification scheme was developed specifically for the FO job. A pragmatic analytical approach and method was adopted by reviewing skills and apportioning them under defined behavioral categories. A unanimous jury procedure was followed in defining the behavioral categories to classify the specifically screened FO tasks. It was inferred that in examining the selected tasks by categories or behavior types, and labelled with terms best describing each task grouping, particular training design relationships would become more apparent and useful in formulating a reliable approach to delineating implications for selection and training.

METHOD

The purpose of the FO task analysis was to identify the essential skills and knowledges that an FO needs in order to perform that combat role. This was accomplished by developing and refining a task list utilizing survey and

[1] This paper is based on research conducted for the US Army Research Institute for the Behavioral and Social Sciences under Contract DAHC-19-78-C-0025 which is being reported under the title: Selection and Training of Field Artillery Forward Observers: Methodologies for Improving Target Acquisition Skills.

[2] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

structured interview techniques and validating that list by obtaining task difficulty and criticality ratings from subject matter specialists (experienced FOs) and Fire Support Team (FIST) Chiefs and by obtaining task ratings from several hundred Field Artillery Officers with operational experience. Once these ratings were obtained, a task selection algorithm was developed which provided a prioritized list of FO tasks. A preliminary list of 118 FO tasks was reviewed by 19 Field Artillery Officer Basic Course (FAOBC) instructors from the Gunnery, Counterfire, and Tactics and Combined Arms Departments at the US Army Field Artillery School, Fort Sill. Nine Gunnery Basic Branch instructors, five Counterfire/Survey instructors and five Tactics and Combined Arms Department instructors were interviewed either individually or in groups of two or three. Based on the specific FO references, direct field exercise observations, and classroom observations, each instructor was asked to verify the completeness of the FO task inventory, to identify any additional tasks that may have been excluded, and to eliminate any non-FO tasks. Additionally, they were asked to comment on the criticality and difficulty levels of tasks relevant to their instructional areas in respect to both the operational and training environments. Also in discussing the impact of task difficulty and criticality on combat and training, the instructors surfaced a problem related to the variability of task differences on several rating dimensions in selected combat scenarios.

Sixty-nine tasks were retained in the final task list upon completion of the instructor review of the preliminary FO task list. Task validation was then conducted on this task list. Fifty-six Field Artillery officers were interviewed, and were assigned as FOs or FIST chiefs attached to operational units, or, had recently performed in the FO role. The participating officers here were from the First Infantry Division (Fort Riley), the 9th Infantry Division (Fort Lewis), and 2d Armor and 1st Cavalry Divisions (Fort Hood). Of the 69 tasks that were included in the FO task list, 44 were rejected as a result of failure to meet the criteria of the task selection algorithm devised and 17 were reinstated by instructor override, with the final FO training list including 42 tasks.

The need to identify tasks that should be included in an FO training program required a decision selection process which led to designing the algorithm which differentially weighted performance, frequency of performance, task difficulty, and criticality ratings. Task analysis summary data and a preliminary selection matrix were discussed with FAOBC instructors with their advisory remarks being equitably applied. If the training time becomes compressed, for example, the minimum cut-off rating scores for criticality and difficulty in the algorithm could be increased which would decrease the number of tasks necessary for training. In effect the tasks selection algorithm can expand or contract to accommodate varying training logistics requirements. There was, then, a task selection scheme applied that attempted to optimize the use of objective criteria in the task selection process.

The 42 training tasks finally specified were divided into five task groups according to the jury procedure and definition mentioned earlier in this paper. These task groups encompassed the following activity types: discrimination tasks, procedure-following tasks, rule-using tasks, problem-solving tasks and tasks involving cognitive/spatial integration skills. While the five task groups may have some partially arbitrary labels, there is sufficiently consistent agreement to noticeably improve the specificity of training task design when using such a technique for task classification after several iterations of improved task identification and definition as described in the several stages of task analysis.

RESULTS

Table 1 describes, for example, the basic structure of the FO task classification scheme which actually shows skill categorization as well. Decisions based on these observed relationships concerning training specificity and the expected level of familiarity can now be more objectively defined. There is a more stimulating choice of what task or skill must have the necessary focus to reinforce the performance required to do the job (Olton, 1979). That is the new FO will be given more flexible task or skill alternatives when adjusting to real changes in the combat or training situations if training is expressed in terms of the applicable task or intercategory sequence.

TABLE 1. *FO TASK SKILL CATEGORIZATION*

Discrimination	*Procedure-Following*	*Rule-Using*	*Problem-Solving*	*Cognitive/Spatial Integration*
Recognize/Identify Target(s)	*Prepare and Transmit a Call-for-Fire*	*Determine Distance by Estimation*	*Prepare and use a Terrain Sketch*	*Orient a map by Terrain Analysis*
	Use MILS as Angular Measurements	*Prepare and Use an Observed Fire Fan*	*Select and Occupy Observation Posts*	*Conduct a Terrain Analysis*
	**Prepare and Transmit a Call-for-Fire (*Context Check)*	*Read a Military Map*		*Navigate on Land by foot*
		Measure an Angle by Using Binoculars		

Discrimination. Only one task was categorized as being primarily a discrimination task and it involved the recognition and identification of targets. Tasks that involve discrimination skills require the individual to determine the differences between or among two or more stimuli and then to respond differently to each stimulus. As Butler (1972) points out, discrimination on a gross level where the differences are clearly defined is a relatively simple task but when the stimuli closely resemble each other, it can be a very difficult task. It is obviously a very critical task for the FO to be able to discriminate enemy targets from friendly forces and materiel. The more varied the number of allied troops on both sides involved in combat, the more complex the task can become. With training time and ammunition already so constrained in the FAOBC program of instruction, discrimination training will probably succeed more when directed toward the more general or gross skill level. The more refined, precise type of target discrimination will find greater improvement as experience is gained and unit level training is effectively disciplined.

Procedure-Following. The second task behavior category was procedure-following which appears to involve the combination of motor and verbal-chaining skills. Procedure-following is the linking together of a series of discriminable responses in a particular order. According to Butler (1972) the recall of the operational procedure becomes dependent upon a chain of responses linked together by both verbal and motor cues. In the FO task categorization scheme, nine tasks fit the description of the procedure-following classification. Because most of these tasks were rated as being very simple to perform in the FO task selection algorithm, many could not meet the composite score criteria for the FO task selection algorithm, and would not have been included in the training task list if they had not been restored to the list by instructor override. It is clear that these tasks are easy to perform and easy to learn but a lot of rehearsal and practice is involved in committing their performance to an almost rote level of competency. This rehearsal and practice can probably be provided best initially in FAOBC since these tasks have a routine application in learning or supporting other associated or higher skill level behaviors. Here and as under the other categories of the task classification scheme, the scope of training becomes a constant issue as it is found that each behavioral category and level of task-skill required will affect training design for other tasks and the training sequence. Intensity-and length-of-training decisions must figure into any training task list and program designed for critical task proficiency.

Rule-Using. The largest group of tasks was the rule-using category which included 24 tasks. This is not really surprising since learning to use rules comprises a large proportion of the specific behavior and knowledge that must be acquired during most types of training. Rule-using behavior requires the individual to learn to perform tasks according to a set of rules or principles. The key to training an individual on rule-using tasks is to provide a number of opportunities in which the rules are methodically applied, not merely state what the rules are. In theory, the FAOBC gunnery field exercises ought to serve this purpose. However, it was usually noted when actually observing "live-firing" exercises that only one student performs the task and the other students observe rather passively if at all. If these exercises, for example, were redesigned so that all students were more involved in the performance of the task, more practice in rule application would perhaps result. One suggestion to illustrate this is for instructors on some "shoots" to have one student indicate the initial target location, another student give the first adjustment, and so on. For this approach to work, however, the instructors would have to maintain constant control to suppress potentially disruptive actions. A second and simple suggestion is to require students to record a location and indicate the adjustments they would make at each step. Even if these step-wise tasks were not graded they could serve to focus student attention on the dynamic firing exercises and would aid instructors in identifying students who may require further assistance.

Problem-Solving. In this category, problem-solving was defined as the ability to solve a novel problem by combining and applying previously learned rules. In the FO training lists, only two tasks met this definition - selecting and occupying an observation post and preparing and using a terrain sketch. Each time the FO is placed into a new setting he must apply the rules that govern these two tasks. Frequent, meaningful practice sessions are the keys to learning these two critical tasks. Sufficient practice is provided in preparing and using a terrain sketch. However, little if any practice is provided in selecting and occupying observation posts. Students are simply taken out to the firing ranges, positioned on the side of a hill and told to adjust fire. An instructor may point out to the class that this is not the way to select and occupy an observation post but never is the student challenged to apply the behaviors for this task. Obviously, then, if students are to be proficient in this task they need to practice it and the "shoot" exercises are the only opportunities for doing so.

Cognitive/Spatial Integration. All of the six tasks included in the cognitive/spatial integration category combine problem solving skills with ability to both convert three-dimensional spatial cues into a two-dimensional projection and analyze the results. These six tasks are the building block tasks of the FO job. If a Field Artillery officer cannot locate the target or himself he cannot adjust fire. Target and self location involve terrain association and map reading skills (Milligan & Waldkoetter, 1979). Interestingly enough, these basic tasks are not taught in FAOBC. Officer students are expected to be able to perform these tasks prior to FAOBC and, consequently, only a quick review is presented by the Counterfire Department at the beginning of OBC. For those students who possess the related terrain association skills and can perform these six tasks, the quick review is sufficient and they are prepared for training as an FO. Those students who do not have these skills begin FO training with a deficit which cannot be compensated for, and they are likely to experience difficulties with many aspects of FO training. There can even be some motivational barrier arising for students who do not perform proficiently on the map reading tests at the beginning of FAOBC. These problems could be minimized by selecting individuals who clearly possess terrain association skills; but a simple solution, it is not practical in this time of personnel shortfalls.

DISCUSSION

There is a tendency to develop task inventories without the process of refinement necessary to make task listings more adaptable to training task formats. Too often task validation becomes the process of administering a long list of specific and overly generalized tasks and then culling the results for what may seem most reliable and descriptive for given jobs and work environments. In this effort to identify and prioritize FO tasks from the larger job performance area down to those crucial training tasks, a four-stage procedure was developed to assure an accurate screening of valid FO tasks. From the initial task listing based on existing references and instructor interviews, a validation by subject matter specialists at several sites, with further administration to several hundred officers for normative data, and the use of the task selection algorithm led to the fourth stage allocation of tasks by the classification scheme.

With 42 training tasks identified and with multistage validation, the allocation under five skill categories led to an economical array of meticuluously selected training tasks. The process of classifying tasks through initial selection, multistage validation and definition of the given categories may seem to be rather detailed but the added precision in the training design and development activities are usually worth this modest investment. Since the five categories and allocated tasks of - Discrimination 1, Procedure-Following 9, Rule-Using 24, Problem-Solving 2, and Cognitive/Spatial Integration 6 - have certain interconnections or nodal relationships, the transference from task to task to trained terminal behavior will occur with more design control. Accordingly, there is then better task selectivity possible relating to more efficient task training decisions which will increase the capability to make more accurate personnel selection or training proficiency inferences.

Although different category structures may be proposed, this training classification and design aid offers a wider determination of just which tasks and task elements may be dominant or dependent. A consistent course of action can be programmed to emphasize given tasks and sequences with clearer paths for diagnosing training problems while prescribing more intensive training procedures. As the FO research unfolded the complex higher-order cognitive tasks seem to have greater dependency on proficiency being achieved in the Rule-Using and Cognitive/Spatial Integration categories. The skills associated with these training task categories could largely affect the proficiency attained by any FAOBC student.

Those FO students who had received extensive training in either science or mathematics where "rule-using" application is practiced more than it would be in a liberal arts program may have developed a rule-use learning strategy. Such a strategy could readily enhance their ability to learn the 24 tasks listed and the key or nodal sequences with other tasks. It is possible that those individuals who have not been trained in the disciplined application of principles may need additional practical exercises and remedial training to achieve certain perceptual criteria. Because the task analysis effort and collection of individual FO background data were performed separately, it was not possible to ascertain directly if individuals with better rule-application skills perform better in FAOBC. However, one finding of the FO background data analysis indicated that those students who were mathematics majors in college performed better in FAOBC. A job sample approach using a simulator, such as the Observed Fire Trainer (OFT) could also serve as a selection device to identify those with the requisite skills. A weapon system training effectiveness analysis study by the Artillery School (USAFAS, 1978) identified the problem of rule-using strategy as a possible source of decreased FO effectiveness, and the tentative solution was to provide the student FO with a training-aid card, listing the rules applicable to a given task. Results from this study indicated no significant differences between the group who had the card and students who did not. Increasing the number of task-oriented exercises in which the student was required to apply the set of rules and improving the quality of the exercises were not addressed. These or related variables are possibly of prime importance in rule-use training.

Under the Cognitive/Spatial Integration Category the six tasks focus on a recurring obstacle in military training. Though selection of personnel demonstrating map reading and terrain association skills is not feasible now, there is every reason to identify students not having these skills and provide additional training at the start of FAOBC. A screening test could be derived from the enlisted course for MOS 13F which trains enlisted FOs and includes a heavy emphasis on terrain association, land navigation and target location tasks. With a rapid expected decay of map-reading skills, those not identified as having these prerequisite training skills may require remedial training to keep up in acquiring FO skills. A dual track (regular & remedial) system might bear some review, but a more proficient level for all tasks and skills could result. Terrain association, land navigation, map-reading, and distance estimation tasks prove very susceptible to variations in task difficulty with different combat scenarios. Little attention is given to these training differences. Due to the compressed training time in FAOBC, the procedural impact of scenario differences may receive only occasional emphasis; however, instructional materials for the unit level could give the necessary practice in adapting to changes in task difficulty and scenario (e.g., Africa).

In developing the task classification scheme a chronic design problem impacting on training arose with an open-ended question requesting experienced FOs to indicate the most important skills for accomplishing the FO tasks. Map reading, fire adjustment, and communication tasks were readily indicated in order and correspond directly with tasks emerging from the task analysis activity. In requesting these tasks, a list of 18 was easily generated with at least three (17%) fitting the label of "soft skills". These tasks are: Understand and work well with maneuver unit; training others; and, leadership. Herein lies a design problem for any task classification scheme which is not based on a firm foundation of accepted definitions. We agreed to regard "soft skills" as those skills emerging from a task allocation process during the design of a job, without specific predesignated action requirements. When asked which tasks FOs performed worst, the same three soft-skill tasks appeared near the top 30 percent of the task list. This simple experimental procedure illustrates that we all nearly miss applying proper classification steps. As we find for task criticality several dimensions to be measured, so also there is a need to perceive the multidimensional need for task classification with each situation or scenario requiring a unique application of occupational analysis methodology. Knowing only too well the arduous pressures in task and training analysis, we must strongly suggest that we shift our perception when classifying tasks to at least view three-dimensional models in our sometimes trial and error analyses. For example, tentatively allow that "leadership" is a category by which "rule-using" tasks can be further analyzed to decide the degree of action or kind applied to the given task category to obtain any specific effects.

We have found the puzzle of task analysis for training stimulating and is resolved best when we can: 1) Identify the requirement; 2) Define the hardware (weapon) and/or software (operator) issues; 3) Determine the component (weapon) and/or skill (operator) to be analyzed; 4) Specify the function (weapon) and/or task (operator) to be performed; and 5) Specify the effects or output needed to satisfy system standards.

REFERENCES

Butler, F.C. Instructional systems development for vocational and technical training. Englewood Cliffs, NJ: Educational Technology Publications, Inc., 1972.

Gagne, R.M. (Ed.), Psychological principles in system development. New York: Holt, Rinehart & Winston, 1962.

Milligan, J.R., and Waldkoetter, R.O. Observer self-location ability and its relationship to cognitive orientation skills (Technical Paper, draft). Fort Sill, OK: U.S. Army Research Institute for the Behavioral and Social Sciences, May 1979.

Olton, D.S. Mazes, maps, and memory. American Psychologist, 1979, 34, 583-596.

U.S. Army Field Artillery School. Weapon system training effectiveness analysis - the forward observer, phase II (ACN 32750). Fort Sill, OK: U.S. Army Field Artillery School, November 1978.

JOB ANALYSIS OF WORK RELATED ATTITUDES[1]

Stanley D. Stephenson

USAF Occupational Measurement Center
Randolph Air Force Base, Texas 78148

In the continuing discussion over the concept of attitudes, a central theme has been the relationship between attitude and behavior. Psychology has generally contended that to know what a person's attitude about a topic is should lead to an accurate prediction of what that person's subsequent behavior will be in relation to that same topic. In fact, if a strong relationship does not exist then perhaps the question might logically be asked, "What is the utility of such a concept?"

At their core, attitudes have been assumed to be a convenient way to summarize some unobservable mental state. If outward behavior can be assumed to be directed by the person's mental state, then obviously the attitude that captures that inward state should also show a similar relationship to behavior as the mental state itself.

However, the relationship between attitudes and behavior has often proven to be inconsistent or even nonexistent. There may be several fairly logical explanations for this lack of relationship (for example, see Wrightman, 1977, pp. 345-348), but the frequently reported attitude-behavior inconsistency has led to a questioning of the value of the concept of attitude itself. Wicker (1969), for example, found little evidence to support a notion of an enduring concept (attitude) that exerts any type of major influence on an individual's verbal or physical behavior, although other authors (Kahle & Berman, 1979; Kelman, 1974; and McGuire, 1976) have produced arguments in support of the attitude concept.

However, knowledge of an attitude is important even if an attitude does not always predict behavior. For example, the attitude, "I don't like this place," may not lead to an employee quitting his/her job, but one might reasonably expect such a person to have a higher probability of quitting than someone not holding that attitude. Such an attitude, if expressed on the job, could also affect the productivity of the holder and others. Moreover, if I am the owner or manager of the company, it would certainly benefit me to know about that attitude for it might be an indicator of a more serious problem within my work force. So, even if attitudes do not always predict behavior, they still have utility. In fact, attitudes are often the only indicator you might have about those factors that are considered important by the holder.

[1] The views expressed in this paper represent those of the author and do not necessarily reflect the views of the United States or the Department of Defense.

There have been two recent attempts to shed more light on this attitude behavior issue. Davey (1976) suggests that attitude-behavior consistency has typically been found in stable social settings in which an accurate behavioral prediction could also have been made based on historical behavioral knowledge alone. He further suggests that perhaps the value of the concept of attitude is in unpredictive situations because a study of such situations could lead to a better understanding of the interaction between the social setting and the individual. Davey argues that it is within the individual-social context relationship that the role of attitude can best be clarified. Such a relationship would include how an attitude is actually formed.

Regan and Fazio (1977) more strongly suggest that the crucial variable is how the attitude was formed. Two individuals holding the same attitude, as measured by traditional attitude scales, may have developed that attitude in completely different ways. Regan and Fazio postulate that all attitudes formed through direct personal interaction with the attitude object are more likely to be predictors of future behavior toward that object than are attitudes formed via some nonpersonal, indirect behavioral contact method. "The person whose attitude is a product of direct interaction with the attitude object will be more likely, in general, to behave consistently with that attitude than someone whose attitude was formed in a less direct manner (p. 31)."

These two approaches are not mutually exclusive, for they both suggest that the method of attitude formation is crucial to understanding the attitude-behavior relationship. If an attitude-behavior relationship is high, it reflects a stable confident disposition. If the relationship is low, then obviously the attitudinal disposition is relatively unstable and/or unconfidently held (for, perhaps, a variety of reasons). In this latter situation the attitude is still an excellent starting point for study because it ties together the relationship (albeit weak) between behavior, the attitude, and the social context that led to the formation of the attitude. Of course, how to determine the method of attitude formation after-the-fact is a major undertaking. Job analysis might be just the tool for this undertaking for it accurately taps what a person does in his or her job. If task performance can be paired with attitudinal responses, the result might prove productive.

Consequently the present report uses job analysis data to investigate the relationship between job related attitudes and the way in which the attitudes were formed. The study specifically hypothesizes that attitudes related to a specific environment will show a stronger task-attitude relationship than will job related attitudes that are not tied to a specific job experience.

Method

Occupational data used in this study were obtained from a job analysis report by Ballentine (1977). The data were collected under the United States Air Force job analysis program described by Christal (1974). This particular job analysis process includes developing an extensive job task list through personal interviews, surveying job incumbents on their performance of the tasks in the list, and analyzing the results using a statistical package (CODAP) especially developed for the USAF job analysis program.

The job analysis reported by Ballentine was conducted on a select group of USAF pilots who had undergone an extensive follow-on training program. The task list contained 197 items organized into eight task categories that covered the non-flying time of the pilots. The survey instrument also contained several open-ended questions that allowed survey respondents to provide additional job information feedback.

Responses to two of these open-ended questions were the starting point for the present study. First, the question was asked, "What did the school not prepare you to do?" A number of the pilots answered, "Manage a training shop". A comparison of task performance was made between these pilots who provided this free response answer (n=32) and those pilots who had not (n=58). Another question asked was, "Why would you leave your present occupation?" Several (n=31) of the respondents answered, "For career broadening," while others did not so respond (n=59). A similar comparison of task performance was also made on these two groups' responses to this question.

These two question were chosen because it appeared that they differed in how they were formed. The -esponse of "They did not train me to manage a training shop" is an attitude based on two distinct behavorial experiences: what was taught at the school and what was done on the job. However, the response of "I would leave my present job for career broadening" is an attitude based on a comparison between a behavioral experience (tasks performed on the job) and a rather abstract perception (I need additional job experience) that may not be based on firsthand knowledge.

The AUTOJET program within the CODAP computer analysis package selects those tasks that show the greatest absolute difference between the average Percent of Group Performing scores of two groups. Tasks are selected for these absolute difference comparisons until a difference of no practical value to a job analyst is reached. If few tasks are selected for comparison, it indicates fairly similar jobs are being performed by the two groups; conversely more tasks indicate more of a job difference.

This type of comparison was conducted for both the Not Prepared to Manage - No Deficiency and for the Career Broadening Needed - Not Needed groups. The nonparametric Sign Test (Sigel, 1956) was used to analyze the results.

Results

For the Not Prepared to Manage - No Deficiency comparison, 80 tasks were selected by the AUTOJET program. All 80 tasks differences were in the same direction; in every case the pilots who had indicated they had difficulty managing their training shop produced a higher Percent of Group Performing average score than did those pilots who had not indicated such a problem. Using the Sign Test, this result is significant at the .001 level.

For the Career Broadening Needed - Not Needed comparison, 27 task differences were selected and analyzed. Ten of the 27 task scores were in the direction of those pilots who felt that they needed to leave their present job for additional job experience. This pattern of Percent of Group Performing differences was not significant as analyzed by the Sign Test.

DISCUSSION

This study confirms the hypothesis that job related attitudes formed in a specific environment have a stronger task-attitude relationship than do job related attitudes that do not have their foundation in a specific environment. However, to know which attitudes were based in a specific environment required job analysis.

It should be noted that the two attitude topics chosen for this study were not random or isolated issues for the individuals involved. For instance, those who believed that they were not prepared to manage also agreed more that they had difficulty managing their shop, that the school did not prepare them to supervise personnel, and that managing a shop prevented them from accomplishing their primary duties. On the other hand those who freely indicated that they needed additional job experience agreed more both that their present job was disadvantageous because of its limited possibilities and also that they wanted to leave their present job to enhance their promotion potential. Obviously both sets of responses were indications of well integrated attitudes.

It should also be noted that in the case of the direct task experience attitudes (the Not Prepared to Manage - No Deficiency comparison) 29 of the 80 tasks selected were in the management and the administrative duty areas; the remaining tasks were spread over the other six duties. Consequently, it could be suggested that the foundation of the attitude, "I was not prepared to manage," lay primarily in those 29 management/administrative tasks. A similar statement can not be made for the career broadening attitudes since a task performance trend did not develop.

Although job analysis of job related attitudes has not been thoroughly researched, support for these results can be found in the literature. For instance, Breer and Locke (1956) stated, "It is our contention that task experience provides much of the raw material out of which men construct their fundamental ideas of life (p. 6)." This study obviously supports that contention. Those pilots who had directly experienced their training school naturally developed the attitude, "the school did not prepare me to manage my shop." In all likelihood these pilots did not have such an attitude before they started their job. The comparable group who did not hold that attitude was not required, for whatever reason, to perform the subset of tasks that led to the attitude development. Performing different tasks should be expected to lead to different attitudes in a direct comparison/experience situation.

On the other hand when a direct task experience comparison was not involved in the attitude development, groups with different attitudes showed no task performance differences. Those pilots who believed they had to leave their present job to broaden their experience were performing essentially the same job as those who did not have this attitude. This can be explained by the need-to-broaden group's perception of their job career world. They may have had no direct contact with the concept of needing more job experience to remain competitive. Their attitude was based on at least second-hand perceptions of the best career route to follow; moreover, they probably held this perception before their present job. In other words attitudes about career development are not entirely dependent on specific tasks performed on a specific job. Groups can, and did, perform similar jobs and still logically differ in their attitudes. Direct task experience - indirect perception is obviously a different situation from direct task experience - direct school experience.

This dichotomy has a direct implication for the study of the attitude-behavior relationship. Since, as Regan and Fazio (1977) suggested, attitudes formed through direct personal interaction with the attitude object are more likely to be predictors of future behavior, knowing which attitudes are formed by direct interaction becomes important. If job analysis can provide this information, then the ability to predict behavior based on attitudes should also increase.

This dichotomy also has a direct implication for attitude change research. In the case where the attitude is a product of direct experience, modifying the experience (e.g., altering the tasks performed or the training) should change the attitude. However, trying to change just the attitude will have little or no effect because the direct experience comparison between tasks and training continues to exist. Conversely, changing the job in the perception based situations should have little effect; in this case the perceptions themselves must be modified because there is no clear behavioral foundation to attack.

Overall, this data indicates that job analysis done in conjunction with an attitudinal survey can increase the understanding of how an attitude was formed. Consequently, job analysis can help predict when an attitude can be expected to lead to future behavior as well as when attitude change can be expected to occur. For a manager dealing with job-related attitudes, such information can be invaluable.

References

Ajzen, I., & Fishbein, M. Attitude - behavior relations: A theoretical analysis and review of empirical results. Psychological Bulletin, 1977, 84, 888-918.

Ballentine, Roger D., Occupational Survey Report of Fighter Weapons Instructors. AFPT 90-5XX-333, USAF Occupational Measurement Center Randolph Air Force Base, Texas, October 1977.

Breer, P.E., & Locke, E.A. Task experience as a source of attitudes. Homewood, Illinois: The Dorsey Press, 1965.

Christal, Raymond E., The United States Air Force Occupational Research Project. AFHRL-TR-73-75. AF Human Resource Laboratory, Brooks Air Force Base, Texas, January 1974.

Davey, A.G. Attitudes and the prediction of social conduct. British Journal of Social and Clinical Psychology, 1979, 15, 11-27.

Kahle, L.R., & Berman, J.J. Attitudes cause behaviors: A cross-logged panel analysis. Journal of Personality and Social Psychology, 1979 37, 315-321.

Kelman, H.C. Attitudes are alive and well and gainfully employed in the sphere of action. American Psychologist, 1974, 29, 310-324.

McGuire, W.J. The concept of attitudes and their relations to behaviors. In H.W. Sinaiko & L.A. Broedling (Eds.), Perspective on attitude assessment: Surveys and their alternatives. Champaign, Illinois: Pendleton, 1976.

Regan, D.T., & Fazio, R. On the consistency between attitudes and behavior: Look to the method of attitude formation. Journal of Experimental Social Psychology, 1977, 13, 28-45.

Siegel, S. Nonparametric methods for the behavorial sciences. New York: McGraw-Hill, 1956.

Wicker, A.W. Attitudes versus action: The relationship of verbal and overt behavioral responses to attitude object, Journal of Social Issues, 1969, 25, 41-78.

Wrightman, L.S. Social psychology (2nd ed.). Monterey, California: Brooks/Cole, 1977.

NOTE

Requests for reprints should be sent to Stanley D. Stephenson, USAF Occupational Measurement Center, Randolph Air Force Base, Texas 78148.

ANALYSIS OF JUNIOR OFFICER TRAINING NEEDS

Richard S. Wellins, Arthur C. F. Gilbert and Michael G. Rumsey

U. S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

INTRODUCTION

The U. S. Army Research Institute (ARI) is currently engaged in a long-term project designed to assist training planners in preparing officers to assume leadership positions. This research effort comes at a time when the Army is seriously examining the officer education and training system. For example, TRADOC's Training and Development Institute is engaged in a comprehensive officer job analysis which will eventually be used as the basis for developing officer training programs. Similarly, ROTC is revising their curriculum and producing a Soldier's Manual for Office Trainees. This manual will contain a list of all subjects and tasks to be taught during ROTC, locations of primary training, and requirements for various tests to insure that the training mission is being accomplished at each detachment and summer camp.

One of the most important training and development phases in an officer's career occurs before commissioning. The Army commissions thousands of new officers annually through three major programs: Officer Candidate School (OCS), the U. S. Military Academy at West Point (USMA), and the Reserve Officers' Training Corps (ROTC). Although the orientation of these programs may differ, they all have the common objective of producing well qualified and highly effective junior officers ready to assume the duties and responsibilities of command. In order to insure that precommissioning training is meeting this objective, it is essential to develop a systematic approach for evaluating and modifying the instructional process.

The research reported in this paper is aimed at the first step in this approach, assessing junior officer training needs. First, the problems that junior officers encounter in trying to accomplish their jobs will be identified and described. Second, suggestions for precommissioning training collected from officers and enlistees in the field will be presented. Lastly, the implications of the research for precommissioning training programs will be discussed. An adequate system of identifying training needs has important uses for curriculum development, both in terms of what should be included in the curriculum and at what level the training should be conducted. In addition, an assessment of junior officer training needs provides insights toward the most advantagous training strategies. This investigation should complement and supplement other officer training analyses by providing a more general picture of the problems encountered by junior officers and how these problems might be resolved through improved precommissioning training.

METHOD

The research was conducted in two phases. First, ARI research teams interviewed over 600 captains, lieutenants, NCOs, and enlistees at seven different U. S. Army installations. Interview participants were asked to elaborate on some of the problems faced by new lieutenants in trying to accomplish their assigned duties. These problems were to be based on the participant's actual experiences. For example, if a participant mentioned discipline as a problem, the interviewer attempted to guide him into relating some of his own experiences in trying to discipline subordinates. After recounting several of these experiences, the participants were given the opportunity to provide recommendations for precommissioning education and training that they thought might better prepare new lieutenants for overcoming some of the problems discussed earlier in the interviews.

Over 60 hours of interviews were on tape by the completion of the seventh field trip. A series of steps taken by the research teams organized and synthesized this massive data base into a form that would be usable by curriculum planners. Listening to the tapes several times made it evident that a number of general problems were mentioned repeatedly during the interviews. It was decided to listen to all of the tapes once again, this time categorizing interview comments under the appropriate problem area. In addition, all training suggestions were listed and then combined into more general categories.

Although patterns emerged from the interviews, it was difficult to quantify the lengthy dialogues contained on the tapes. For example, although several of the participants complained about being overburdened with secondary duties, it was impossible to determine from the interview data alone how important this problem really was in relation to other frequently mentioned problems, or how it impacted on overall job effectiveness. Consequently, a questionnaire designed to quantify and validate the information gathered during the interviews was administered to a sample of 228 officers, NCOs, and enlistees.

The major portion of the questionnaire was devoted to validating the problem areas identified during the interviews. Officers were requested to rate these problem areas, using five-point scales, on three dimensions: the extent to which they personally experienced the problem, how widespread they thought the problem was among their peers, and how important the problem was in terms of job performance and leadership effectiveness. Enlistees were asked to rate the problem areas on only the "widespread" and "importance" dimensions. In addition to the problem-area ratings, several suggestions for training that had been mentioned repeatedly during the interviews were restated as questionnaire items. Respondents were asked to indicate how valuable they thought each of these suggestions would be in terms of preparing new officers to handle their jobs more effectively.

RESULTS

The critical problem areas identified in the interviews and validated in the questionnaire will be discussed here. Those training areas which were emphasized during the interviews and rated as very valuable on the questionnaire will also be presented. The detailed statistical analyses of the data will be part of a later ARI Technical Report.

Problem Areas

After leaving Officer Basic Course, new lieutenants are thrust into leadership positions in which they must assume immediate command. Yet, many of the officers we interviewed complained that they were insufficiently trained in the leadership and management techniques essential to effectively lead their troops. The training they did receive was often criticized for being irrelevant or unrealistic. For example, one officer expressed a concern that training in leadership theory was a waste of time unless the principles of the theory could be related directly to the job. Similarily, new officers frequently lacked the skills to manage resources, people, and time. In the field, the lieutenant is required to be a competent military leader; in garrison, the lieutenant's job is more like that of a mid-level manager in a large corporation. He must be trained accordingly.

Closely related to the command and leadership area are the problems a junior officer encounters in applying military law and in disciplining subordinates. Many of the comments on military law centered around the handling of drug problems. What does a second lieutenant do, after being in his unit only a month, when he suspects or actually finds drug use among his subordinates? Does he go by the book, let them off, or make some sort of compromise? Improper handling of a drug problem can put the new officer in an embarrassing situation as well as decrease troop morale. Disciplining troops in today's Army also presents a difficult situation for the typical lieutenant fresh out of college. As one participant stated, "You can't go out and give a guy an order anymore and expect it to be obeyed." The interviewed lieutenants complained of low quality personnel who could not be depended upon. Often, the soldier does just enough to get by and will not follow an order until he is actually threatened with an Article 15 or some similar action. Such formal discipline procedures often take too long to initiate and are frequently rejected higher up in the chain of command, even further frustrating the lieutenant.

Counseling is another important aspect of the new lieutenant's job which requires interpersonal skills. One officer estimated that over 50% of a second lieutenant's time is spent in a counseling capacity. Comments gathered during the interviews and the results of the survey, however, indicated that most cadets receive only rudimentary training in the skills and knowledges necessary to effectively counsel subordinates. New lieutenants are frequently confronted with subordinates' personal problems ranging from marital difficulty to letters of indebtedness. In addition, the officer is expected to counsel subordinates on job performance. No matter what purpose counseling serves, it seems critical that new lieutenants receive appropriate training in this area before they encounter a situation which they are unable to properly handle.

One of the most persistent problems mentioned during the interviews and validated by the questionnaire was the misunderstanding of the role and utilization of the NCO. One officer stated that, "the most important problem for a new lieutenant when he first comes on duty is understanding the relationship and value of the NCO." The NCOs interviewed agreed that a good relationship with the NCO is imperative for successful officer performance. Frequently, however, interpersonal and organizational barriers obstruct good relations. For example, the new lieutenant will often walk into his unit with a "know-it-all" attitude, telling the NCO what to do and how to do it. Obviously, a 40-year-old NCO with 15 years of service experience is likely to resent a "kid" fresh out of college telling him what to do. The NCO frequently reacts to this attitude by taking a passive role and allowing the lieutenant to fail. On other occasions, however, the NCO will take a more active role and actually sabotage and new officer by making him look bad. Other officers alleged that incompetent NCOs interfered with unit effectiveness. Apparently the Vietnam war led to the premature promotion of many NCOs who did not yet have the experience needed to carry out their assigned duties.

The inability to effectively communicate with subordinates poses yet another problem for the new officer. The majority of new lieutenants come from a college environment where most of their interactions are with peers. Even their military training offers little opportunity to work with the NCOs and enlistees they will soon be leading. Consequently, lieutenants often arrive on the job with a limited understanding of the cultural, social, and economic background of their subordinates that may prevent effective communication. It was evident from the interviews that officers were frustrated because their subordinates were unable to understand the simplest orders. At the same time, enlistees complained that officers were constantly talking over their heads, using vocabulary they didn't understand.

Although most of the problems encountered by new officers were interpersonal in nature, several of the participants complained that lieutenants were inadequately trained in some technical or hard skills. It is difficult the teach new officers every hard skill they will need to know prior to arrival at their units; some of their skills will have to be learned on the job. Nevertheless, several of the new officers felt a great deal of pressure to perform every task competently or risk a bad Officer Efficiency Rating from their superiors. Other participants felt that skill deficiencies could interfere with establishing effective interpersonal relationships with the enlistees and NCOs in the unit. If the officer is not properly trained in the skills he needs, the men in his unit may try to take advantage of the officer. The largest number of complaints on job skill deficiencies came from officers who felt they were not prepared to deal with the load of secondary duties that they were typically assigned. They also felt that many of these secondary duties detracted from combat effectiveness and the time they could take to become familiar with the men in their units. Other participants readily gave examples of officers who were not adequately trained in their primary duties or in even more basic military skills such as land navigation and map reading.

Training Suggestions

Participants made several insightful suggestions for improving precommissioning training which were later validated by the questionnaire data. The single most frequently mentioned training suggestions was that all cadets go through the Cadet Troop Leadership Training (CTLT) program. The objective of this program is to provide cadets with realistic leadership experiences while assigned to units of the active Army. The cadets are given some of the responsibilities of a second lieutenant. Currently, all West Point cadets go through the program while only a small number of ROTC scholarship students have this opportunity. Although CTLT was repeatedly mentioned as a valuable training vehicle by participants of all ranks, some of these participants cautioned that the program must allow the cadet a certain amount of freedom and responsibility. If the cadet is treated as a trainee or assistant and given minor duties, the experience will be wasted.

A second suggestion was the use of NCOs for training ROTC cadets. Currently, NCOs are used to provide instruction in certain technical areas. However, the interviewed participants suggested that NCOs might provide cadets with a different perspective of Army life as well as give them advice on how to establish a proper and effective Officer-NCO working relationship.

Other suggestions centered around improving training programs in areas such as management, counseling and leadership. The lieutenants and captains we interviewed complained that current training in these areas was often unrealistic and based more on theory than practice. Training in soft skills should be more related to the actual requirements of the job, allowing cadets the opportunity to practice some of the skills necessary for effective performance in these areas.

DISCUSSION

The findings of this investigation tended to coalesce into groupings of fairly consistent themes. Perhaps the most prevalent theme was the importance of emphasizing soft skills in precommissioning training. Soft skills might be generally described as skills used to accomplish tasks which are not defined in clear, precise, detailed, terms. Thus, such tasks require, rather than the technical abilities to adapt a specified set of procedures to a particular problem, the officer's ability both to define the nature problem and establish the framework of the solution.

Many of the problems identified in the soft skill areas were of an interpersonal and organizational nature. The officers interviewed reported difficulty interacting with their NCOs and communicating with their subordinates. The implications of these problems for training became more important when participants in the research project were asked to make suggestions for precommissioning training that would better prepare junior officers for their first positions. The suggestions consistently stressed the importance of giving ROTC students experience in interacting with others in a military environment prior to being commissioned. Insofar as this experience gives the cadets a better understanding of the background and needs of the enlistee and the organizational relationship with the NCOs, it should better prepare the cadet for interacting with these individuals.

The feedback obtained in this investigation also revealed problems in the soft skill area of command and leadership, including unfamiliarity with the Army system and inadequate leadership and management training. Counseling and discipline problems also required a greater degree of soft skill expertise than many officers felt they had. In order to prepare ROTC cadets to handle these problems, respondents stressed the importance of training which would provide experiences in dealing with realistic, job-related problems rather than merely providing theory.

A second theme that emerged from this investigation is that officers seemed to feel that they were relative well training in hard skill areas. Discussion of problem areas generally focused on soft skill problems. Two major exceptions to this trend were in the areas of secondary duties and basic soldiering skills. However, even though hard skill areas generally seemed to produce fewer problems than soft skill tasks, there was no indication from this study that officers felt hard skills were unimportant. On the contrary, it was suggested that competence in hard skills gave an officer greater credibility with the troops and thus enhanced that officer's capability to lead.

The results of this research project will provide valuable information to those responsible for designing precommissioning curriculum. In fact, some of the findings from this study have already had some impact on current TRADOC efforts to standardize instruction for ROTC. The first draft of the Soldier's Manual for Officer Training (SMOT), a standard list of core subjects/tasks to be trained in ROTC, did not include many of the soft skills identified in this investigation as critical to effective junior officer performance. Plans are now underway to expand the SMOT to include some of these areas. The result should be the production of more effective junior officers ready to assume the duties and responsibilities of command.

POLICY SPECIFYING WITH APPLICATION TO PERSONNEL CLASSIFICATION AND ASSIGNMENT

Joe H. Ward, Jr., Manuel Pina, Jr.,
Jonathan C. Fast, and David K. Roberts

Air Force Human Resources Laboratory (AFHRL)
Brooks AFB Texas 78235

INTRODUCTION

Personnel of the Air Force Human Resources Laboratory have been developing and applying procedures for creating models of judgment processes for over twenty years. The initial efforts were concerned with representing the judgments of Air Force personnel classification specialists. Out of these efforts have grown many insights into Judgment Modeling, as well as new ways to look at this problem. This paper discusses the latest development at AFHRL in Judgment Modeling and details the evolution of a new technique known as Hierarchical Policy Specifying. An example of how this technique has been used in personnel classification and assignment will be presented.

OBJECTIVE AND EXAMPLES

The objective of Judgment Modeling is to combine several different types of information into a single indicator of "value," "payoff," or "utility." A significant operational example of this is the Air Force Weighted Airman Promotion System (WAPS) which is used for promotion to grades E-4 through E-7. The objective of this project was to determine how to combine several different types of personal information, such as time in grade, specialty knowledge, awards and decorations, and performance ratings, into a single index of promotability. In this case, a technique known as judgment analysis (or policy capturing) was used to develop the weighting equation. Another example within the Air Force was the need to combine information about people and jobs to reflect the "payoff" to the Air Force of assigning a particular person to a particular job at a particular time. This was accomplished using a technique known as policy specifying within the Air Force Person Job Match (PJM) system. Other examples of the need to combine multiple information into a single value are: graduate admissions policy, performance appraisal, financial analysis, research and development resource allocation, officer grade requirements, job difficulties, and national recovery policy.

JUDGMENT MODEL DEVELOPMENT

The Air Force has developed some methodologies for dealing with these types of situations using systematic approaches. The procedures that will be presented in this paper are not meant to exhaust the list of ways to deal with procedures for combining multiple items of information to produce a single representation of the payoff judgments of one or more individuals. Some other methods are listed in the bibliography which include scaling methods (Saaty, 1977), behavioral decision theory (Slovic, Fischhoff, and Lichtenstein, 1977) and utility theory (Keeney, and Raiffa, 1976). Three methods that will be examined in this paper and have been used in the Air Force are:

THREE METHODS

- POINT ALLOCATION
- JUDGMENT ANALYSIS or POLICY CAPTURING
- POLICY SPECIFYING

POINT ALLOCATION

The first method will be called point allocation. This method requires the identification of the relevant variables to be weighted and the desired payoff range. The expert consulted for the procedure allocates various percentages of points to each variable to be combined to yield the total payoff.

POINT ALLOCATION

TOTAL PAY-OFF RANGE = 100

Percent Pay-Off Allocated to X1 = 20

Percent Pay-Off Allocated to X2 = 30

Percent Pay-Off Allocated to X3 = 50

Can yield an explicit equation

$$Y = W_1 X_1 + W_2 X_2 + W_3 X_3$$

Given the percentage allocation points, an explicit equation could be defined, but in practice, the values of the W's and X's are not usually determined.

JUDGMENT ANALYSIS

The term judgment analysis will be used in this paper to describe the situation where judges are given decision situations and then their judgments are analyzed. The following steps are used in the judgment analysis procedure:

JUDGMENT ANALYSIS

Step 1. Identify the concept of interest (Y) and the variables (X's) which will be used to describe Y.

Step 2. Select one or more judges and a sample of decision situations to be judged.

Step 3. Judges assign values of Y to each decision situation.

Step 4. Obtain least squares regression to predict judgments in the form

$$Y = w_1 X_1 + w_2 X_2 + \ldots + w_p X_p + E$$

An example of how judgment analysis is used will be presented, using as the judge a young airman entering the Air Force for his first assignment. The judgment situation will be various assignment locations within the Air Force that the airman must rate. The example will consist of the following three variables:

JUDGMENT ANALYSIS EXAMPLE

Preference for Assignment Location (Y)

Function of:

- Cost of Living (X1)
- Population of Community (X2)
- Annual Snowfall (X3)

The airman is presented 150 samples of Air Force Base locations. These locations are not described in terms of name or location, but in the following terms:

PRESENT JUDGMENT SITUATIONS

Sample of 150 Air Force Base locations described in terms of

	Range
• Relative Cost of Living (X1)	1 (low)-7 (high)
• Population of Community in Thousands (X2)	0 - 1,000
• Annual Snowfall in Inches (X3)	0 - 100

Judgments are then obtained from the judge for each of the samples. He expresses his preference for each location using values from -9 (low) to +9 (high). For purposes of this example, the airman will be assumed to be a single, 19-year-old male from rural Minnesota. From the expressed preference values, the least squares weights are computed and a regression equation formed to predict his preferences. An example might be:

COMPUTE LEAST SQUARES

WEIGHTS TO PREDICT PREFERENCES

$$Y = W_0 + W_1 X_1 + W_2 X_2 + W_3 X_3 + E$$

$$Y = (-9) + (-.00003) X_1 + (-.0002) X_2 + (.18) X_3 + E$$

This equation could be interpreted to show that the airman was most concerned with variable X_3, annual snowfall, since he was from Minnesota. The airman was least concerned about cost of living in this example since he was single. This equation could be used to predict the airman's preference for another location that he had not previously judged.

HIERARCHICAL POLICY SPECIFYING

The term Hierarchical Policy Specifying will be used in this paper to describe a decision theory methodology developed within AFHRL. This technique consists of the following steps:

POLICY SPECIFYING

Step 1. Identify the fuzzy concept of interest (Y) and the variables (X's) which will be used to describe Y.

Step 2. Select one or more Experts who will specify policy in mathematical form.

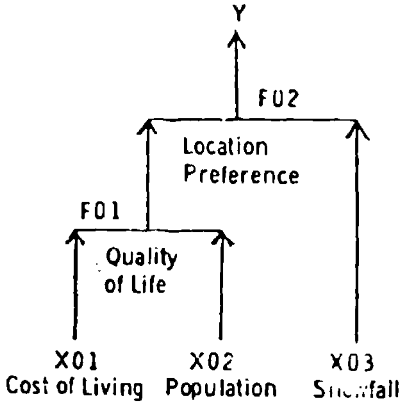
Step 3. Define a pair-wise hierarchy of the X's.

Step 4. Specify the pair-wise models.

The first step is similar to the first step of judgment analysis, except that often the concept of interest (Y) is fuzzy, i.e., there exists no device with which to measure the payoff or value of the alternatives available to the judge. Examples of this are the values to the Air Force of promoting an individual, and of classifying an individual in a certain job. Step 3 is probably the most important step and possibly the most difficult step in the policy specifying procedure. The judge must interface with the modeler to define a hierarchy for the variables (X's) which will be used to specify the policy. An example of the pair-wise policy is constructed for the previous example of the pre'erence for assignment location.

POLICY SPECIFYING EXAMPLE

PAIR-WISE HIERARCHY

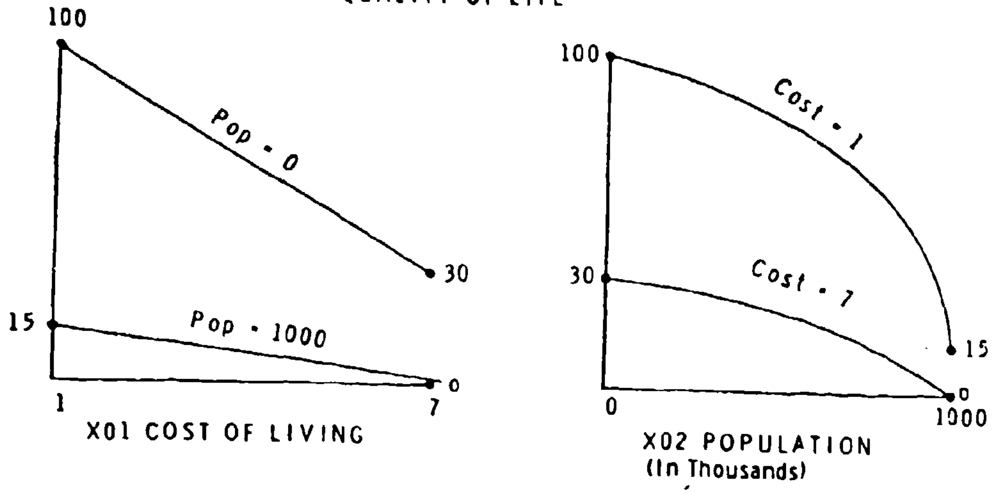


In this example, the judge and the policy modelers have decided that the variables X01 and X02 should be combined first using function F01. The output of this function will be called Quality of Life. Then the output of function F01 will be combined in the hierarchy with Snowfall (X03) using function F02 to give Y (Location Preference). As can be seen from the example, the order of combination in the hierarchy is somewhat arbitrary and different results might occur with a different order in the hierarchy. This is a research question which must be pursued in the future before too much can be said about the stability of a policy specified in this way.

Given the hierarchy, the modeler must then elicit from the judge responses which will allow him to construct functions F01 and F02. The expert is asked to describe how he feels about Cost of Living and Population. He specifies his preferences at the four extreme combinations of the two variables. In this example, the four values specified are 100, 30, 15, and 0. Then the judge specifies how the function output varies for each of the variables. In this case, he indicates that for different values of Cost of Living, the function output change is constant. This indicates the linear relationship shown. However, for the Population variable, he indicates that he is relatively indifferent for low values of Population, but as the population becomes large, the change in function output is much greater. This indicates the second order non-linear function shown.

FUNCTION F01

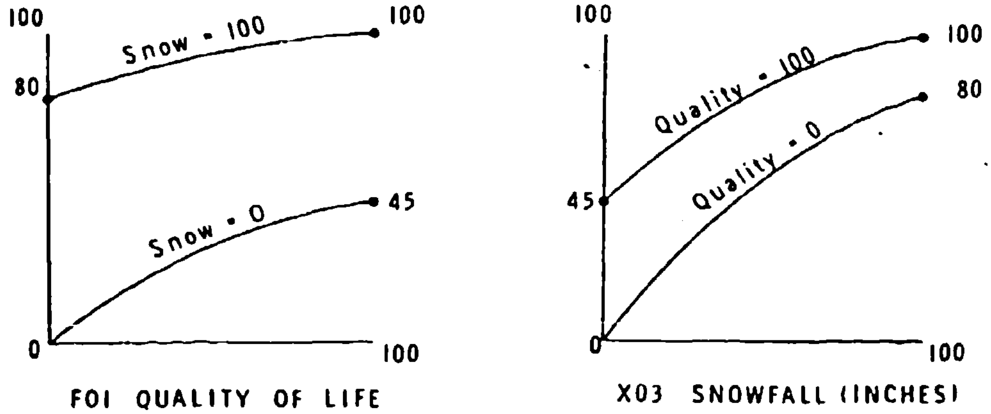
QUALITY OF LIFE



The judge then proceeds to describe function F02, relating the output of F01 to Snowfall. For this function, the judge specifies the four extreme values as 100, 80, 45, and 0. He then specifies that for different values of Quality of Life, the function output varies as a non-linear curve of second degree. However, for Snowfall, he specifies that for low values of Snowfall any changes in Snowfall make a big difference, but for high values, he is relatively indifferent. The model maker translates this specification into a third order relation since the distinction between low and high values of Snowfall seems stronger then for Quality of Life.

FUNCTION F02

LOCATION PREFERENCE



The equation generated from the policy capturing procedure and the expanded version of the final policy specified equation (which is usually not explicitly developed) are as follows:

COMPARISON OF MODELS

Judgment Analysis

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + E$$

Policy Specifying

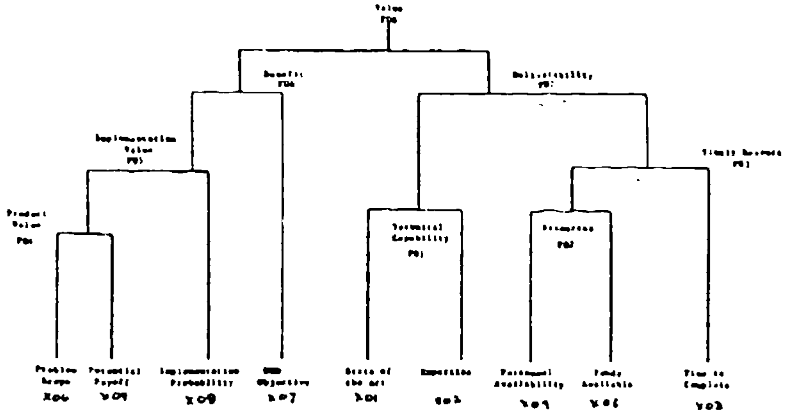
$$Y = a_0 + a_1 X_1 + a_2 X_2^2 + a_3 X_3 + a_4 X_1^2 + a_5 X_1 X_2^2$$

$$+ \; . \; . \; . \; +$$

$$a_{33} X_1^2 X_2^4 X_3^3 + a_{34} X_2^2 X_3^3 + a_{35} X_2^4 X_3^3$$

The judgment analysis model generally contains no product terms (interaction) and no higher order terms. Note also that the judgment analysis model contains an error term that can be used to determine accuracy of predicting the judgments. Policy specifying has been used for other applications within AFHRL. This is an example of a hierarchy specified for research evaluation and investment decisions.

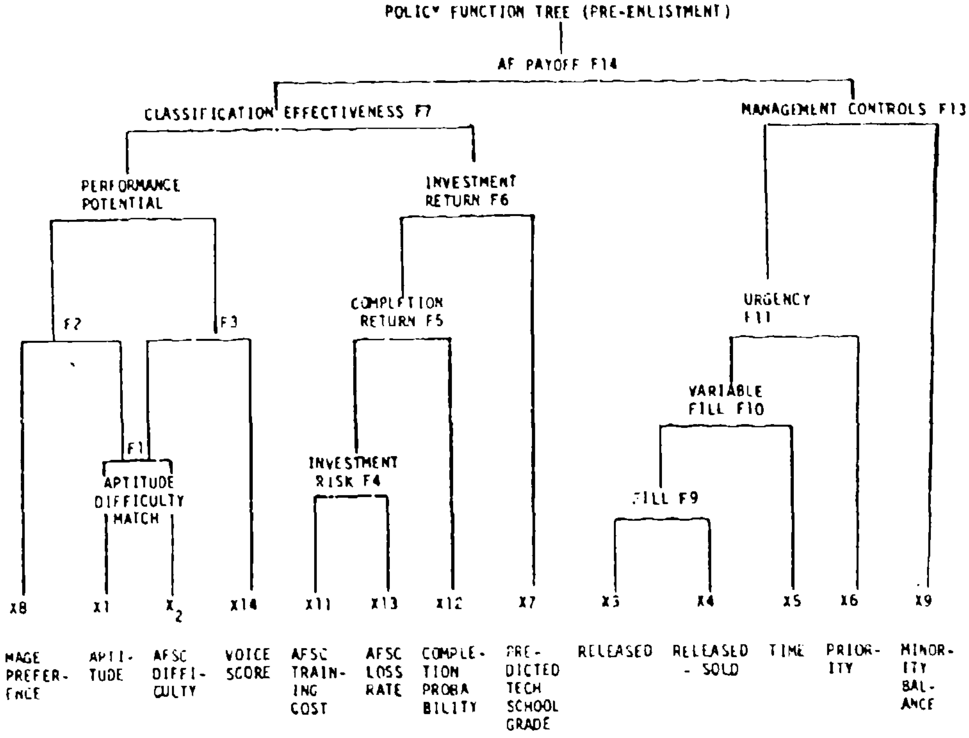
R&D RESOURCE ALLOCATION HIERARCHY



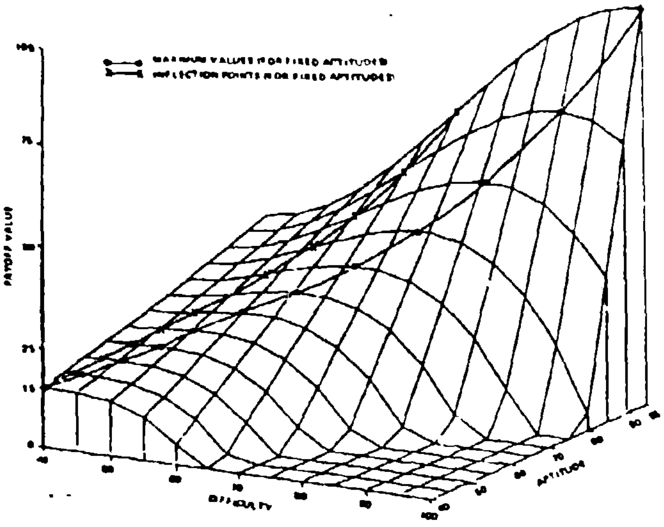
HIERARCHICAL INTERACTION OF INPUT VARIABLES

RESEARCH EVALUATION & INVESTMENT DECISION SYSTEM (REIDS)

In another example, policy specifying has been used to define a hierarchy for the initial classification of Air Force recruits.



In this hierarchy, the experts in conjunction with the model makers at AFHRL, have defined the variables which are important in determining the payoff to the Air Force of classifying each airman into each AFSC. The experts then specified the hierarchy of functions that combines the variables in a pair-wise fashion. The individual's aptitude is matched with the difficulty of the job, using function F1. The surface shown below relates aptitude to difficulty. Then the airman's preference for an AFSC is matched with F1 to form Performance Potential. When the Vocational Interest Career Examination (VOICE) is available in the AFEES, this score will be substituted for Mechanical, Administrative, General, or Electronics (MAGE) preference to form Performance Potential (using F3 in place of F2). The second portion of the hierarchy yields a function output called Investment Return (F6) and as can be seen from the variables input to it, this hierarchy represents the amount of return on the invested dollar that management can expect to get from an airman. The variables at the lowest level of this hierarchy are AFSC Training Cost, which is the cost of being trained into this AFSC, and the Loss Rate for that AFSC. These two are combined to form an Investment Risk function (F4), with a high risk value for any given AFSC representing either a large investment or a high loss rate. This function is then combined with the applicant's probability of completing the first term of enlistment as calculated using the Likelihood Function Estimation (LIFE)



Pay-Off Function of Aptitude and Difficulty.

Model. The output of this function, Completion Return (F5), tends to give a higher payoff for an individual with a high Completion Probability in a high-cost/high-risk AFSC, and a low payoff to the low Completion Probability individual. In the hierarchy, Performance Potential and Investment Return are combined by function F7 to yield Classification Effectiveness. The last function, F14, provides the overall payoff to Air Force by combining Classification Effectiveness with Management Controls.

SUMMARY

The primary distinctions between the three methods of decision theory modeling discussed here are:

SUMMARY OF METHODS

- POINT ALLOCATION
  - Attempts to produce a policy model by allocating percentages of total points to each of the X's.
- JUDGMENT ANALYSIS
  - Attempts to predict behavior of a judge by computing a weighted function of the X's.
- POLICY SPECIFYING
  - Attempts to produce a policy model by specifying a functional relation among the X's.

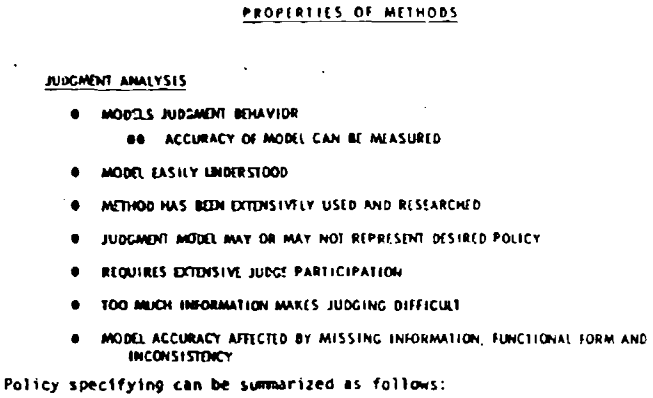
Point Allocation has the following properties:

PROPERTIES OF METHODS

POINT ALLOCATION

- EASY TO USE AND EASILY UNDERSTOOD
- MAY BE ADEQUATE FOR SOME POLICIES
- DIFFICULT TO EXPRESS COMPLEX POLICIES
  - DOES NOT USE INTERACTIONS OR NON-LINEARITIES
- MAY NOT MODEL JUDGMENT BEHAVIOR

Judgment analysis has these properties:



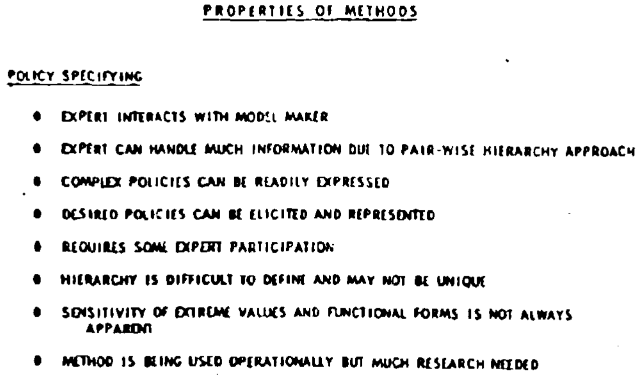
PROPERTIES OF METHODS

JUDGMENT ANALYSIS

- MODELS JUDGMENT BEHAVIOR
  - ACCURACY OF MODEL CAN BE MEASURED
- MODEL EASILY UNDERSTOOD
- METHOD HAS BEEN EXTENSIVELY USED AND RESEARCHED
- JUDGMENT MODEL MAY OR MAY NOT REPRESENT DESIRED POLICY
- REQUIRES EXTENSIVE JUDGE PARTICIPATION
- TOO MUCH INFORMATION MAKES JUDGING DIFFICULT
- MODEL ACCURACY AFFECTED BY MISSING INFORMATION, FUNCTIONAL FORM AND INCONSISTENCY

Policy specifying can be summarized as follows:

Policy specifying can be summarized as follows:



PROPERTIES OF METHODS

POLICY SPECIFYING

- EXPERT INTERACTS WITH MODEL MAKER
- EXPERT CAN HANDLE MUCH INFORMATION DUE TO PAIR-WISE HIERARCHY APPROACH
- COMPLEX POLICIES CAN BE READILY EXPRESSED
- DESIRED POLICIES CAN BE ELICITED AND REPRESENTED
- REQUIRES SOME EXPERT PARTICIPATION
- HIERARCHY IS DIFFICULT TO DEFINE AND MAY NOT BE UNIQUE
- SENSITIVITY OF EXTREME VALUES AND FUNCTIONAL FORMS IS NOT ALWAYS APPARENT
- METHOD IS BEING USED OPERATIONALLY BUT MUCH RESEARCH NEEDED

Hierarchical Policy Specifying shows much promise, but it also requires a significant amount of research to better define the method. Alternate procedures for specifying the functions and sensitivity analysis of the specified functions are two research projects which will be started in the near future. In the meantime, other applications for the specifying methodology are being pursued, including an advanced enlisted assignment system and an officer initial classification system.

BIBLIOGRAPHY

Black, D.E. Development of the E-2 weighted airman promotion system. AFHRL-TR-73-3, AD-767 195. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, April 1973.

Christal, R.E. JAN: A technique for analyzing group judgment. The Journal of Experimental Education, Summer 1968, 36(4), 24-29. (a)

Christal, R.E. Selecting a harem - and other applications of the policy-capturing model. The Journal of Experimental Education, Summer 1968, 36(4), 35-41. (b)

Dempsey, J.R., Sellman, W.S., & Fast, J.C. Generalized approach for predicting a dichotomous criterion. AFHRL-TR-78-84. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, February 1979.

Gooch, L.L. Policy capturing with local models: The application of the AID technique in modeling judgment. Unpublished doctoral dissertation, The University of Texas at Austin, 1972.

Gott, C.D. Development of the weighted airman screening system for the air reserve forces. AFHRL-TR-74-18, AD-781 747. Lackland AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, March 1974.

Jones, K.M, Mannis, L.S., Martin, L.R., Summers, J.L., & Wagner, G.R. Judgment modeling for effective policy and decision making. Research Report for Air Force Office of Scientific Research Grant No. AFOSR-74-2658, AD-A033 186.

Keeney, R.L., & Raiffa, H. Decisions With Multiple Objectives: Preferences and Tradeoffs. New York, N.Y.: John Wiley & Sons, 1976.

Koplyay, J.B. Extension of the weighted airman promotion system to grades E-8 and E-9. AFHRL-TR-70-2, AD-703 687. Lackland AFB, TX. Personnel Research Division, Air Force Human REsources Laboratory, January 1970.

Koplyay, J.B., Albert, W.G., & Black, D.E. Development of a senior NCO promotion system. AFHRL-TR-76-48, AD-A030 607. Lackland AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, July 1976.

Mullins, C.J., & Usdin, E. Estimation of validity in the absence of a criterion. AFHRL-TR-70-36, AD-716 809. Lackland AFB, TX: Personnel Division, Air Force Human Resources Laboratory, October 1970.

Saaty, T.L. A Scaling Method for Priorities in Hierarchical Structures. The Journal of Mathematical Psychology, 1977, 473(15), 111-158.

Slovic, P., Fischhoff, B, & Lichtenstein, S. Behavioral Decision Theory. The Annual Review of Psychology, 1977, 28, 1-39.

Ward, J.H., Jr., & Davis, K. Teaching a digital computer to assist in making decisions. PRL-TDR-63-16, AD-407 322. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, June 1963.

Ward, J.H., Jr., & Jennings, E. Introduction to linear models. Englewood Cliffs, NJ: Prentice-Hall, 1973.

Ward, J.H., Jr., & Haltman, H.P. Computer-based enlistment quota reservation system using the general data management system 2000: Programming and implementation details. AFHRL-TR-75-71, AD-A021 340. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human REsources Laboratory, December 1975.

Ward, J.H., Jr. Creating mathematical models of judgment processes: From policy capturing to policy specifying. AFHRL-TR-77-47, AD-A048 983. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, August 1977.

POLICY SPECIFYING, JUDGMI T ANALYSIS, AND NAVY PERSONNEL ASSIGNMENT PROCEDURES

Dr. Leonard Kroeker

Navy Personnel Research and Development Center
San Diego, California 92152

A paper presented at the Annual Military Testing Association Couference. San Diego, California, 17 October 1979.

Introduction

The problem being addressed is that of improving the match between recruit applicants and entry level Navy ratings. From the standpoint of recruiters and classifiers, it makes sense to discuss a Navy job for which the applicant is suited. A mis-match can be costly, for both recruiting efforts in the short run, and for retention efforts in the long term. Both of these points will be elaborated.

The existing PRIDE reservation system requires a judgment concerning where a recruit applicant might best serve the Navy. The new model will assist the recruiter, classifier, and applicant to reach such an assignment decision. In fact, the model will ensure that desired Navy manpower policy is substantially represented in these decisions.

The CLASP model is in the form of a computer program and is designed to be completely compatible with the PRIDE reservation program. When the two have been integrated, the computer terminals currently in use may be used to effect both classification and reservation functions.

CLASP is an easy way to use procedure, capable of handling first term non prior service male applicants. It is written in the form of an interactive computer terminal dialogue prompting the user to enter data concerning:

a. biographical information;

b. aptitude test scores;

c. shipping dates preferred; and

d. personal job preferences.

Within 15 to 30 seconds an ordered school seat list is printed at the terminal and specific Navy ratings on the list, are discussed with the prospect. The ratings are ordered so that the ones at the top of the list best meet the joint needs of the individual and the Navy.

Problem Definition

The person-job matching problem may be conceptualized as an interaction between an individual and an organizational system. We think of effective recruiting as taking place at the interface between these two systems. Person P is represented by certain attributes and the Navy is represented by job J's attributes. To the extent that the two sets of attributes mesh, a good assignment results, with the resulting benefits of increased job satisfaction and personnel retention.

CLASP reflects this conceptualization in that it allows the assignment decision to be decomposed into separate utility components. This allows differing individual and organizational objectives to be quantified and integrated. It also permits trade-offs among potentially conflicting objectives to occur.

The Utility Model

In the assignment process a number of aspects concerning the person-job match are considered. Some of these are more germane to the individual and others are more important to the organization within whose domain the job lies.

In many instances the individual may evaluate the job on a set of dimensions relevant for him but different in part from the set of dimensions used by the organization to evaluate him. The totality of dimensions consisting of individual and organizational dimensions may be considered as a common space characteristic of this particular person-job match. Clearly, the match must involve a degree of compromise on a number of the dimensions on the part of either the individual or the organization.

The utility model described is a first approximation and is concerned with both Navy classifier behavior and judgment and that of the recruit applicant. The modelling of the assignment process therefore reflects both the individual's and the organization's objectives.

Navy classifiers hold very definite opinions about classification decisions they make but are much less definite about describing the process underlying such a decision. Under careful scrutiny, it is apparent that different classifiers depend upon different sets of salient decision variables or cues. This is not an unexpected result in view of the fact that they need not hold identical belief and value structures. In fact, it is not uncommon that a classifier changes the implicit weighting of the variables or the constituency of the variable set even though similar circumstances would call for consistent application for the previously followed decision procedure. One of the purposes of CLASP is to formalize the classification decision procedure to ensure consistent application.

The Navy has traditionally based classification decisions largely upon paper and pencil tests that have been validated against final school grades. One of the problems encountered in restricting attention to these tests is that direct measures of an individual's ability to perform the required tasks within a given job are not directly assessed. A comprehensive classification model should incorporate components that depend upon such measures.

The proposed assignment procedure is based on an additive linear model consisting of components that play an integral role in the decision process resulting in the placement of an individual in a given job. It incorporates the notion of payoff or utility. A larger payoff value is more desirable because the probability of an individual succeeding in a job is a monotonically increasing function of increasing payoff value.

The classification procedure operates on a payoff matrix, which is a rectangular array of numbers that represent the utilities of decision outcome combinations involving the assignment of a given individual to a particular job. The utilities express the value to the Navy (on an arbitrary scale) of the consequences of a particular decision.

The utility of assigning person P to job J is derived from a weighted linear combination of utilities each of which expresses a different aspect of the decision outcome. The five utility generating components are identified as:

1. The school success component;

2. the technical aptitude/job difficulty component;

3. the Navy priority/individual preference component;

4. the minority fill-rate component; and

5. the fraction fill-rate component.

The School Success Component

School success is defined as final A-school performance grade and is used as the criterion variable in linear regression involving ASVAB subtest scores as predictor variables. For each of the 86 ratings a regression equation is developed such that the squared value of the multiple correlation between the criterion and predictor variables is at the maximum.

The Technical Aptitude/Job Complexity Component

In the process of ascertaining whether or not an individual is appropriately suited to a particular job an employer must establish both the nature of the job, its tasks and requirements and the collection of abilities a person possesses taken as a global construct. In other words, does this particular person possess the set of abilities required for success on the job.

During a typical employment interview an employer makes a judgment with respect to a given individual using some internal scale. The scale is usually not well defined but resembles an unidimensional continuum fashioned from a collection of ability scales considered jointly. It is effective however in that it allows the employer to order prospective employees along a continuum. He might well come up with a judgment such as, "John Doe appears to belong to the upper 25% of the applicants when assessed on the internal aptitude scale."

The job on the other hand is somewhat more familiar to the employer and he is certainly able to delineate the characteristics of the job and the type of individual most likely to fill the job successfully. In fact, he is able to describe a job in terms of the technical ability it requires. This enables him to order jobs on an unidimensional continuum on the basis of the technical ability required and it forms a second scale. Let us suppose that the job in question has been rated by the employer as belonging to the upper 25% of jobs with respect to the above-mentioned criterion.

After having established both the relative position of the job on the respective scales it is necessary to make a judgment concerning the correspondence of the individual and the job. In this case there appears to be a good match and it is likely that the individual will be offered the job.

The aptitude/difficulty component in the utility model acts in a manner similar to that of the employer described above. It attempts to establish a correspondence between individuals and ratings such that a more able person is considered for a more demanding job and vice versa.

The Navy Priority/Individual Preference Component

It is clear that considerations involving both Navy priorities and individual preferences are essential in establishing an appropriate correspondence between a prospective recruit applicant and a rating. In other words the person-job match affected should reflect the serious consideration of individual preference whenever possible and also the fulfillment of Navy priority requirements.

The two sets of objectives may not be compatible with one another, particularly if both are described by utility functions that are allowed to vary independently. For example, the gain in utility resulting from an expression of strong preference for a given rating may be off-set by a loss in utility stemming from the fact that the rating in question may hold a very low Navy priority value.

To overcome the deficiency of a strictly additive model an interactive utility function was designed in order to capture the policy interventions of military decision makers. It yields an utility value as a function of the Navy priority index value for a particular rating in conjunction with the specified preference value of an individual for the rating.

The Minority Fill-Rate Component

For a number of reasons minority group members have been assigned in disproportionately large numbers to a few ratings and in disproportionately small numbers to many others. One of the problems that the CLASP is to address is the one concerning adequate representation of minority group members in all ratings.

A strict quota which involves the reservation of a fixed number of positions for minority group members within each of the 86 ratings appears to be an unpalatable solution in view of recent legal proceedings. The utility model approach, on the other hand, has several advantages to recommend it.

First, it avoids the inflexibility that may prove harmful to both the individual and the Navy when an unsatisfactory person-job match occurs under quota system constraints. Secondly, it allows the component charged with accomplishing the desired fill proportions within ratings to influence the assignment process only to the extent desired by the military decision maker who must be sensitive to dynamic operational constraints.

The underlying idea of the minority fill-rate component is that of a uniform rate of acquisition of minority group members within a given rating. Of course a uniform rate on Non-Minority group member acquisition is also implied. Specifically it is desired that the proportion of minority group members within any particular rating at a given time always equals the previously specified goal of a specific minority group proportion for the rating.

The difference between the actual and desired minority group proportions at any given time may be used as an indicator of the status of the uniform rate of fill objective and may be employed as the driving mechanism of a feedback function. The function compensates for existing conditions by either awarding additional utility points when the actual proportion is less than that desired or subtracting utility points otherwise.

The Fraction Fill-Rate Component

Under current operating conditions the end of any given recruiting month is marked by a flurry of recruiting activity aimed at filling a substantial number of positions within certain ratings. This situation is brought about largely by the fact that no systematic mechanism exists to influence the assignment procedure to exhibit uniform acquisition rates within ratings.

From a managerial perspective a procedure resulting in a uniform rate of acquisition is highly desirable. The Fraction Fill-Rate Component has been designed to meet this objective.

The compnent compares the proportion of applicants assigned to positions within a given rating with the average proportion of applicants assigned to all ratings at any given time. If the proportion for a given rating is below the average value, additional utility points are awarded in order to influence the applicant to select the rating. If the rating is selected, the proportion corresponding to the rating then moves closer to the average value.

In a similar fashion, utility points are removed when the proportion of applicants assigned to openings within a rating exceeds the average value. If the rating in question is not selected and another rating is selected instead, the resulting average proportion value increases slightly thereby moving toward the value of the proportion corresponding to the rating under consideration.

The Fraction Fill-Rate Component therefore acts as a feedback mechanism and is driven by the discrepancy between the proportion for a specific rating and the average proportion. The advantage of this type of component lies in the fact that it is not directly time dependent. Instead of being strictly time dependent it relies on the feedback characteristic to accomplish a uniform rate of fill across ratings. It is important to note that the procedure does not pre-suppose a uniform rate of fill across time but rather accomodates any prospective applicant arrival rate that one may encounter.

The Optimal-Sequential Assignment Procedure

The brief account found below describes the assignment process and draws material from the discussion in the previous sections.

A recruit applicant, interested in a Navy career, completes the ASVAB test battery, submits to a physical examination, and provides demographic and personal data, all of which is used to ensure that he will be placed in a position whose requirements are commensurate with his ability level.

Initially, the applicant is considered as a potential candidate for every rating, and for each rating he receives a composite utility score. It is obtained by calculating and weighting the value of the five utility components. The composite utility score is compared with the average composite utility value for the given rating and the difference is called the decision index score. Thus, an applicant has a decision index score for every rating.

The decision index scores for a given individual are then scaled. The resulting scores have certain useful properties and are referred to as optimality index scores. The scores enable the ratings to be rank ordered from highest to lowest, where the highest ranked rating is the most optimal one for the individual since relative utility is greatest. The list is reduced by removing all ratings for which the individual does not qualify or for which there is no opening within the recruiting month in question.

The amended list is then presented by the classifier to the applicant with the most optimal jobs on the top of the list being discussed first. Under ideal circumstances, the applicant always selects the first rating listed since it is associated with the maximum utility when individual and organizational objectives are considered jointly. In practice, however, one expects an individual to select a rating further down the list since he is evaluating ratings solely from his own perspective as opposed to a joint organizational and individual perspective. Therefore, the applicant has a tendency to examine and select ratings from further down the list whereas the classifier has a tendency to focus upon and emphasize the ratings at the top. When a compromise is reached a reservation is made for the rating agreed upon.

INTERACTION AMONG PEOPLE CHARACTERISTICS AND JOB PROPERTIES IN DIFFERENTIAL CLASSIFICATION

Joe H. Ward, Jr.

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

INTRODUCTION

A major responsibility of personnel systems developers is to provide personnel managers with techniques to select, classify, assign and reassign people to appropriate jobs that tend to maximize overall system effectiveness. These techniques involve both the definition and measurement of relevant attributes of people (referred to as person characteristics) and the definition and measurement of relevant attributes of jobs (referred to as job properties). The various items of information about people and jobs then are combined to yield an indicator of expected future performance (referred to as "payoff", "value", "worth", or "utility") of each person on every job (Ward, 1977). Using these indicators of payoff, personnel are assigned to jobs in a way that will tend to maximize the overall system effectiveness (usually by maximizing the sum of the person-job payoff values associated with the designated assignments). While techniques might be available for designating without choice which person will go into which job (Langley, Kennington, Shetty, 1974), most personnel assignment systems allow for some but not total freedom of choice by both personnel (for jobs) and by job managers (for personnel). This leads to a requirement for techniques that provide for each person-job assignment combination an indicator of the desirability of that particular assignment for overall maximization of personnel assignment effectiveness. Such an indicator has been referred to by the general term Allocation Index (Ward, 1978). However, the Allocation Index may be called by other more appropriate names such as Decision Index (DI), Differential Assignment Index (DAI), Differential Classification Index (DCI), Optimality Index (OPI), or Personnel Optimality Index (POPI). An Allocation Index is dependent on the particular group of people and particular set of jobs being considered in the assignment system. However, the Payoff Value for a person-job combination is generally independent of other people and other jobs being considered for assignments. The term Optimality Index (OPI) will be used in this discussion to refer to an indicator of overall personnel systems effectiveness.

An Optimality Index (OPI) that has been used by the Air Force is a Decision Index. (Ward, 1959) This index was developed to provide helpful information to Air Force personnel specialists to assist in making good personnel classifications.

Recently, an important relation has been recognized between the problem of optimal personnel assignment (and its related Optimality Indicator) and the concept of interaction[1] among people and jobs.

This recognition opens up a wide range of new opportunities to improve the effectiveness of personnel assignment systems. Several of these are:

1. A measure is available for the extent to which it is possible to make any improvement in personnel systems effectiveness.

2. Assuming that there is a big potential for assignment effectiveness (as indicated by a large amount of interaction), it is possible to measure for each job the importance of making a good personnel assignment to that job.

3. It is possible to measure for each person the importance of making a good job assignment for that person.

4. It might be possible to use a set of person measures or job measures in a non-interacting form in association with other predictor information used in interactive ways on only a limited sample of data to obtain a more accurate prediction system. Then the non-interactive measures need not be collected for future operational predictions. This feature has potential for reducing the amount of operational testing required or may eliminate the need for operational use of controversial predictor variables.

Before discussing in detail the four implications mentioned above, a general overview of personnel assignments and the equivalence of the assignment problem and interaction will be presented.

A GENERAL VIEW OF PERSONNEL ASSIGNMENTS

A general view of an array of Predicted Payoff Values is shown in Figure 1. There are Np "real" people to be considered for assignments to a total of Nj jobs consisting of a total N1 jobs of Type 1 plus N2 Jobs of Type 2, through N_L jobs of Type L. To allow for the possibility of all jobs being unfilled (i.e., occupied by a Shadow Person) and to allow for all persons being rejected from the personnel system (i.e., assigned to External Jobs), we allow for a total of Nj Shadow Persons and Np External Jobs. Therefore, there is a total of N (=Np (Real) + Nj (Shadow)) Persons to be considered for N (=Nj (Internal)) + Np (External) Jobs.

[1] Interaction refers to the "constant difference" question studied in Analysis of Variance and the General Linear Model. (Bottenberg and Ward, 1963; Ward and Jennings, 1973)

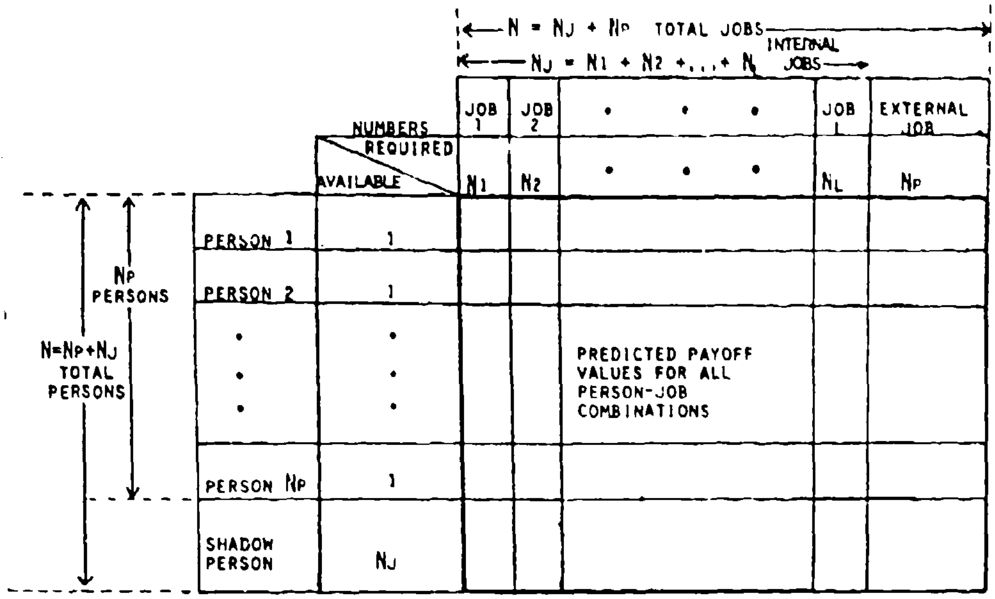
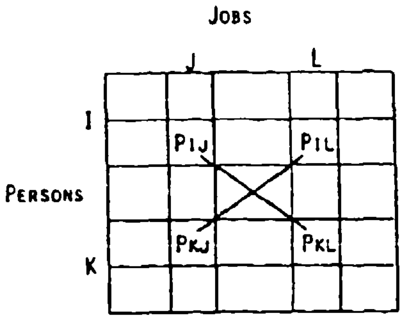


FIGURE 1. PREDICTED PAYOFF ARRAY

For the remainder of the discussion it is assumed that the Predicted Payoff Array is of dimension N by N. There are therefore N! possible allocations of N persons assigned to N Jobs. For each allocation a total payoff can be calculated from the N values associated with each person-job assignment. Many of these different assignments might produce the same sum. The N! values are referred to as the Objective Function, Z.

OPTIMUM ASSIGNMENTS AND INTERACTION

The problem of Optimum Assignments is to assign persons so that the sum of the payoff values (objective function Z) is maximum. If the elements of the Predicted Payoff Array are designated P, then the concern is to compare alternative assignments as shown in Figure 2.

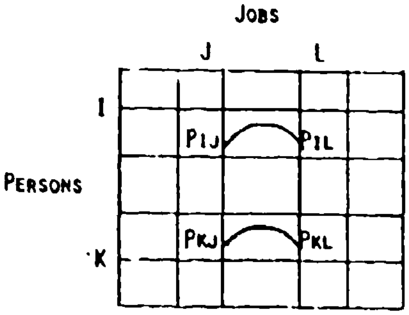


$$(P_{IJ}+P_{KL}) - (P_{IL}+P_{KJ}) = (CS)_{IJKL}$$

FIGURE 2. OPTIMUM ASSIGNMENTS

If the comparisons (CS)ijkl (for all possible persons and jobs) are equal to zero, then all of the N! assignments will give the same objective value. And of course the larger the values (CS)ijkl then different assignments can make a bigger difference in the objective function.

Now notice Figure 3 and a conceptually different view. The interest here is in applying analysis of variance or a general linear model approach to the study of interaction among people and jobs.



$$(P_{IJ}-P_{IL})-(P_{KJ}-P_{KL}) = (CD)_{IJKL}$$

FIGURE 3. INTERACTION AMONG PEOPLE AND JOBS

In this case if all comparisons (CD)ijkl are equal to zero then it is said that no interaction exists. And, as above, larger values of (CD)ijkl indicate that more interaction is present.

$$(P_{IJ}+P_{KL})-(P_{IL}+P_{KJ})=(P_{IJ}-P_{IL})-(P_{KJ}-P_{KL})$$

$$(CS)_{IJKL}=(CD)_{IJKL}$$

FIGURE 4. OPTIMUM ASSIGNMENTS EQUALS INTERACTION

Figure 4 summarizes the fact that the basic concerns of the investigation of optimum assignments are equivalent to the concerns about interaction among people and jobs.

1. Potential Assignment Improvement From Variance of Objective Function, Optimality Index, And Interaction

To obtain an indicator of the potential for improved assignments, we examine the relation between the variance of the Objective Function, variance of the Optimality Index, and interaction.

The interaction sum of squares for the N by N payoff array is shown in Figure 5.

$$S = \sum_{I=1}^{N} \sum_{J=1}^{N} (P_{IJ})^2 - \frac{1}{N} [\sum_{I=1}^{N} (\sum_{J=1}^{N} P_{IJ})^2 + \sum_{J=1}^{N} (\sum_{I=1}^{N} P_{IJ})^2] + \frac{1}{N^2} [\sum_{I=1}^{N} \sum_{J=1}^{N} P_{IJ}]^2$$

FIGURE 5. INTERACTION SUM OF SQUARES

Figure 6 shows the Mean and Variance of the N! values of the objective function, Z.

$$\underline{\text{MEAN}} \qquad\qquad \underline{\text{VARIANCE}}$$

$$\bar{Z} = \frac{1}{N}\left[\sum_{I} \sum_{J} P_{IJ}\right] \qquad \sigma_Z^2 = \frac{S}{(N-1)}$$

FIGURE 6. MEAN AND VARIANCE OF Z

And Figure 7 defines an Optimality Index (OPI) for person i on job j and gives the mean and variance of the N^2 values of OPI.

<u>OPTIMALITY INDEX</u>

$$(OPI)_{IJ} = P_{IJ} - \frac{1}{N} \sum_{I} P_{IJ} - \frac{1}{N} \sum_{J} P_{IJ} + \frac{1}{N^2} \sum_{I} \sum_{J} P_{IJ}$$

$$\underline{\text{MEAN}} \qquad\qquad \underline{\text{VARIANCE}}$$

$$\overline{(OPI)} = 0 \qquad \sigma_{OPI}^2 = \frac{S}{N^2}$$

FIGURE 7. MEAN AND VARIANCE OF OPI

The value of (OPI)ij is a direct indicator of the extent to which the assignment of person i to job j can be expected to increase (if positive) or decrease (if negative) the overall objective function. OPI values can be computed and used to determine an ordered job list from which a person may choose a job. And, in addition, the OPI values can be used to determine an ordered personnel list from which a job manager may select.

Figure 8 summarizes the relation between the variance of the Objective Function, variance of the Optimality Index, and the interaction sum of squares.

$$\sigma_Z^2 = \frac{S}{N-1}(\text{INTERACTION SUM OF SQUARES})$$

$$\sigma_{OPI}^2 = \frac{S}{N^2}(\text{INTERACTION SUM OF SQUARES})$$

FIGURE 8. VARIANCE OF Z, OPI, AND INTERACTION

It is now apparent that as the interaction approaches zero, the variances of the Objective Function (Z) and the Optimality Index (OPI) approach zero. These three indicators measure the potential improvement for making better personnel assignments.

2. Job-Potential Index as Measured By The Variance of The Job's Optimality Indexes

If the variance of the OPI values is large (i.e., there is a large amount of interaction among people and jobs), it is of interest to determine that part of total interaction variance that is associated with each job. The variance of the OPI values for the job c gives this measure as shown in Figure 9.

$$\sigma_{(OPI)c}^2 = \frac{1}{N}\sum_{i}[P_{ic}]^2 - \frac{2}{N^2}\sum_{i}P_{ic}[\sum_{j}P_{ij})$$
$$+\frac{1}{N^3}\sum_{i}[\sum_{j}P_{ij}]^2 - \frac{1}{N^2}[\sum_{i}P_{ic}]^2$$
$$+\frac{2}{N^3}[\sum_{i}P_{ic}][\sum_{i}\sum_{j}P_{ij}] - \frac{1}{N^4}[\sum_{i}\sum_{j}P_{ij}]^2$$

FIGURE 9. VARIANCE OF OPI FOR JOB C

If a job has a very large $\sigma^2_{(OPI)c}$ compared with other jobs' variances, then it is very important to attend to the task of making a "better assignment" for that job. Otherwise, that job has a good change of receiving a bad personnel assignment. A proposed name for this measure is the Job-Potential Index.

3. Person-Potential Index For Each Person as Measured By The Variance of Each Person's Optimality Indexes

Having considered the variance of the OPI for each job, it is also of interest to compute a similar measure for each person. The variance of the OPI for each person is shown in Figure 10.

$$\sigma^2_{(OPI)R} = \frac{1}{N}\sum_{J}[P_{RJ}]^2 - \frac{2}{N^2}\sum_{J}P_{RJ}[\sum_{I}P_{IJ}]$$
$$+\frac{1}{N^3}\sum_{J}[\sum_{I}P_{IJ}]^2 - \frac{1}{N^2}[\sum_{J}P_{RJ}]^2$$
$$+\frac{2}{N^3}[\sum_{J}P_{RJ}][\sum_{I}\sum_{J}P_{IJ}] - \frac{1}{N^4}[\sum_{I}\sum_{J}P_{IJ}]^2$$

FIGURE 10. VARIANCE OF OPI FOR PERSON R

Similar to the discussion of the Job OPI variance, if a person has a very large $\sigma^2_{(OPI)r}$ compared with other persons' variances, then it is very important to attend to the task of making a better assignment for that person. A proposed name for this measure is the Person-Potential Index.

4. Improving Prediction And Reducing The Number of Predictors

The relative quality of an allocation of persons to jobs is not affected by adding (or subtracting) a constant to an entire row (or column) of the Predicted Payoff Array. This opens the interesting prospect of improving the accuracy of prediction of the payoff values by including among the predictor variables a subset which do not involve any person-job interaction. These predictor variables are referred to as the Non-Interactive Variables. The other predictor variables are called Interactive Variables. The prediction system is developed using both the Interactive and the Non-Interactive Variables, thereby obtaining the regression weights for the Interactive Variables in the presence of the Non-Interactive ones. Then the operational assignment system applies only the regression weights for the Interactive Variables (computed in the model with the Non-Interactives) to the values of the Interactive Variables--the Non-Interactive Variables are not required.

If improved prediction can be achieved in this manner, then fewer predictor variables are required. This reduction in variables could reduce testing time and/or allow for the elimination of controversial variables.

SUMMARY

Recognition of the relation between optimal assignment procedures and interaction among persons and jobs suggests consideration of the recommendations shown in Figure 11.

1. USE AN OPTIMALITY INDEX (OPI) IN OPERATIONAL ASSIGNMENT SYSTEMS.

2. COMPUTE FOR A PREDICTED PAYOFF ARRAY EITHER:

   A. VARIANCE OF THE OBJECTIVE FUNCTION
   B. VARIANCE OF THE OPTIMALTY INDEX
   C. INTERACTION SUM OF SQUARES,

   AND USE ONE OR MORE OF THESE AS A MEASURE OF POTENTIAL FOR ASSIGNMENT IMPROVEMENT.

3. COMPUTE FOR EACH JOB THE VARIANCE OF THE OPTIMALITY INDEX AND USE THIS VARIANCE AS A MEASURE OF THE IMPORTANCE OF MAKING A GOOD ASSIGNMENT TO THAT JOB.

4. COMPUTE FOR EACH PERSON THE VARIANCE OF THE OPTIMALITY INDEX AND USE THIS VARIANCE AS A MEASURE OF THE IMPORTANCE OF ASSIGNING THE PERSON TO A GOOD JOB.

5. EXPLORE THE POSSIBILITY OF REDUCING THE NUMBER OF PREDICTORS AND IMPROVING PREDICTED PAYOFFS REQUIRED IN PERSONNEL ASSIGNMENT SYSTEMS.

FIGURE 11. SUMMARY OF STEPS

BIBLIOGRAPHY

Bottenberg, Robert A., & Ward, J.H., Jr. Applied multiple linear regression. PRL-TDR-63-6, AD-413 128. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, March 1963.

Hendrix, W.H., Ward, J.H., Jr., Pina, M., Jr., & Haney, D.L. Pre-enlistment person-job match system. AFHRL-79-29, Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, September 1979.

Langley, R.W., Kennington, J., & Shetty, C.M. Efficient computational devices for the capacitated transportation problem. Naval Research Logistics Quarterly, Office of Naval Research, December, 1974, v. 21, No. 4, 637-647.

Pina, M., Jr. & Stifle, J.L. Person-job match computer-based research system. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, in press.

Ward, J.H., Jr. Use of a decision index in assigning air force personnel. WADC-TN-59-38, AD-214 600. Lackland AFB, TX: Personnel Laboratory, Wright Air Development Center, Air Research and Development Command, April 1959.

Ward, J.H., Jr., & Davis, K. Teaching a digital computer to assist in making decisions. PRL-TDR-63-16, AD-407 322. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, June 1963.

Ward, J.H., Jr., & Jennings, E. *Introduction to linear models*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

Ward, J.H., Jr., & Haltman, H.P. *Computer-based enlistment quota reservation system using the general data management system 2000: Programming and implementation details*. AFHRL-TR-75-71, AD-021 340. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, December 1975.

Ward, J.H., Jr. *Creating mathematical models of judgment processes: From policy-capturing to policy-specifying*. AFHRL-TR-77-47, An-A048 983. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, August 1977.

Ward, J.H., Jr., Haney, D.L., Hendrix, W.H., & Pina, M., Jr. *Assignment procedures in the Air Force procurement management information system*. AFHRL-TR-78-30, AD-A056 531. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, July 1978.

PERFORMANCE vs. PAPER-AND-PENCIL ESTIMATES OF COGNITIVE ABILITIES

James K. Arima

Naval Postgraduate School
Monterey, California 93940

INTRODUCTION

The use of traditional, psychometrically created, paper-and-pencil tests for selection has come under considerable criticism in recent times. One dominant source of this critical appraisal is equal employment opportunity legislation and the court decisions that have followed. The tests have been criticized for their cultural bias, and even when they have been shown to be equally valid for various ethnic or socioeconomic groups in the job context, their continued use has been decried on the basis of the adverse impact that results. Another source of criticism has been politically motivrted actions capitalizing on the distrust and dislike of objective tests by a segment of the general public. This activity has resulted in the banning of mass testing for pupil classification in California and the so-called "truth in testing" legislation passed in New York (Smith, 1979). Finally, questioning of the construct validity and ecological relevance of factorially developed tests has come from the lack of intersection between test constructs and findings in the rapidly developing area of cognitive psychology (Carroll & Maxwell, 1979; Sternberg, 1979). This last basis for criticism is particularly important to the psychological profession as it points out the distinction made years ago by Cronbach (1957) of the two disciplines of scientific psychology--the correlation and the experimental approaches.

Taking cognizance of these trends, an earlier effort attempted to create a performance test that was practical to administer, had high construct validity, was culture free, and would provide results that could be broadly generalized (Arima, 1978; Young, 1975). In addition, an important consideration in creating the test was to measure an ability that was not being sufficiently assessed by conventional testing procedures and that would simultaneously provide a new dimension for making selection decisions. Accomplishing this could increase the selection pool and provide opportunities for individuals who might have been eliminated by conventional procedures. The new dimension was learning aptitude, defined as the ability to profit from experience. Broadly defined in this manner, learning ability has been proposed as an important indicator of intelligence and that higher levels of intelligence would be demonstrated by the ability to learn a fixed amount of material in a shorter time or a larger amount of material in a fixed period of time (Estes, 1974). Learning ability, manifested by such measures as grade point average, has been frequently used as a dependent variable in traditional test research, but the format and procedures of paper-and-pencil tests have made it impractical to use learning as an independent or selection variable. On the other hand, simple learning tasks have been extensively used and validated in comparative psychology (Bitterman, 1975). Validation in this context has been the demonstration of reliably different levels of performance in humans by age or in animals by the phylogenetic hierarchy (Jensen, 1979).

The test, itself, was a discrimination-learning task in which pairs of random forms were presented sequentially to subjects. One member of a pair was arbitrarily designated as the correct alternative, which the subject learned to identify on the basis of positive reinforcement whenever a correct choice was made. Six different pairs made up a list, and their presentation, a trial. In all, 10 trials were given with the item pairs appearing in different random orders in each of the trials. The test was administered in a machine-paced and a self-paced model to Navy recruits undergoing basic training.

Significant amounts of learning took place over the 10 trials, and the correlation between odd and even trials showed a reliability of .838 when corrected for a test of full length using the Spearman-Brown formula. There was a low, but significant correlation (r = .27, N = 137) between the discrimination-learning test scores and the Armed Forces Qualification Test (AFQT) scores attained by the subjects in their entrance testing. When the total group was split into white and nonwhite subjects, only the correlation for the white subjects (r = .223, N = 104) reached statistical significance at the .05 level. Thus, it appeared that the performance measure might be giving an assessment of the true capability of the nonwhite subjects which the verbal AFQT score failed to accomplish. Since, however, the correlation was .213 for the 33 nonwhite subjects, its lack of significance might have been due to smaller sample size. There was also a significant difference on the learning test between white and nonwhite subjects using the machine-paced mode, but not in the self-paced mode. However, the interaction term of ethnic grouping and presentation mode had a probability between .10 and .20 in the analysis of variance of learning test scores, so the differential effects of presentation mode for the racial groupings was not fully confirmed.

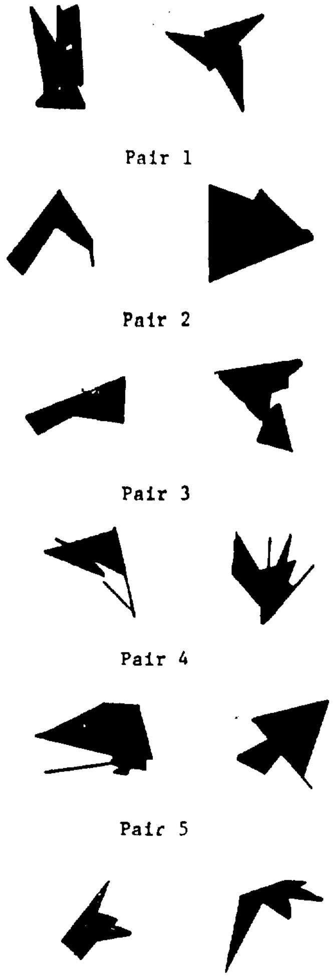
The present effort was a continuation of the original project that was motivated by several reasons. First, the learning test was reconfigured to make it more portable and simple to administer. It was made into a self-paced mode using a correction procedure so that selection of only the correct alternative automatically advanced the test to the next pair of items. These changes required a tryout and comparison of the results with the previous findings. There was a desire to see if the lack of a difference in performance between whites and nonwhites would hold up in the self-paced mode using the reconfigured test. There was also a severe restriction in range in the earlier study because the subjects had been selected for service using the AFQT score as a screen. An unselected group was desired for whom the scores of the entire selection battery would be available for comparison with the discrimination-learning test score.

METHOD

Test Modifications

The test, as developed for the original study (Arima, 1978), had three stimulus "lists" that were presented to individuals and scored by means of a set of "off the shelf" laboratory equipment. It was basically a machine-paced test, and the subject had to press a button to advance the stimuli in the self-paced mode. The equipment was cumbersome and large and required considerable effort to set up. The objectives of the test modifications were to make it simple and portable and to run automatically in a self-paced mode.

Since there was no great effort for similarity of stimuli within or between the lists in the original study, stimulus list 1 from the original study was selected. This list (Fig. 1) was constructed to have the least amount of similarity between the stimuli in each pair and among the pairs of the list. One member of each pair was randomly designated as the correct choice.



Pair 1

Pair 2

Pair 3

Pair 4

Pair 5

Pair 6

Figure 1. Test Figures

Table 1

Subjects

Ethnic Group	Male	Female	Total
White	17	26	43
Nonwhite	11	11	22
(Total)	28	37	65

The basic equipment for the reconfigured test was an SR-400 Stimulus-Response (S-R) Programmer made by Behavioral Controls, Inc. (BCI). The SR-400 has four clear-plastic panels that can be used to present visual stimuli and also serve as the response keys. Stimuli are presented by means of a fan-folded continuous strip of paper that can be programmed to control each of the four channels. It is essentially a sophisticated "teaching machine." In this application, only the two central panels were used, and the other two were blacked out and deactivated.

As previously, 10 different versions of the stimulus list were made in which the order of the pairs was different, and each member of a stimulus pair randomly occupied the right or left position an equal number of times over all 10 versions. The 10 lists were connected into one continuous sequence with the restriction that any one pair did not appear back-to-back. The lists were physically created by pasting the appropriate random figures to the designated position (right or left) on a sheet of the continuous, fan-folded paper. Each pair was coded for the correct response by punching the appropriate channel of the control segment of the sheet. This was done for the 60 stimulus pairs that constituted the entire, 10-list sequence.

In operation, the SR-400 was programmed to advance to the next stimulus pair when the correct panel (stimulus) had been pressed. Thus, a correction method was used for the reinforcement--i.e., the subject had to make a correct response before the paper would move. A BSI counter incorporated into the setup through a BCI Four-Choice Auxiliary Control Console cumulated correct and incorrect responses, and a timer mounted on the control console cumulated viewing time. (It did not move during the time the programmer was cycling to a new pair.) A stepping counter was built into the rear of the counter to buzz when six consecutive correct responses were made, but it became unreliable and was not used in test runs. The cycle time between stimulus pairs was 1.4 sec., and the equipment was programmed to stop at the end of the 10-list sequence.

Subjects

Subjects were obtained through three high schools in Monterey County, California, that participated in the high school testing program of the Defense Department. In this program, the Armed Services Vocational Aptitude Battery (ASVAB) is administered as a service without cost to high schools for vocational counseling. The results of the testing go initially to the high school counselor, but copies also go to recruiters in the area of the participating schools. Utilizing this source of subjects made it possible to compare learning performance with psychometric test measures in a relatively unselected population, which was one of the purposes of this study. The 65 students with ASVAB scores who were made available for this effort were divided by sex and ethnic grouping as shown in Table 1. The nonwhites were Hispanic (11), black (1), Filipino (2), Oriental (4), Native American (1), and other (3). The subjects came from grades 9 through 12 with the average being 10.7. They ranged in age from 14 through 18 with an average age of 16.2 years.

ASVAB

The ASVAB used in the high school testing program was the version identified as Form 5. The tests of the battery, along with their length and reliability, are shown in Table 2. The General Information test includes items of common knowledge that individuals could pick up casually. It was included to provide a measure of the ability of subjects who do not do well in the remainder of the battery, especially those coming from socially deprived environments. Attention to Detail (AD), a perceptual speed test, and Numerical Operations are designed to evaluate potential clerical workers. The Electronic (EI), Shop (SI), and Automotive Information (AI) tests are trade-type tests to identify individuals who already have some capability in these areas or whose familiarity with the material serves as an indication of their interest in this type of work. The other tests are assessments of cognitive skills and stored knowledge. The Armed Forces Qualification Test (AFQT) score is a linear combination of the Word Knowledge (WK), Arithmetic Reasoning (AR), and Space Perception (SP) tests normed on the World War II mobilization population. It has a reliability of .93 (Jensen, et al., 1977). The utilization of the ASVAB in high schools for counseling has been criticized by Cronbach (1979) because it is essentially a selection and placement test as used by the Armed Forces. The Armed Forces Vocational Testing Group has attempted to create composites and provide norms using the relevant population to make it more acceptable for counseling in the high schools while still retaining its primary purpose for the military (U.S. Military Enlisted Processing Command, undated).

Procedure

The test equipment, now quite portable, was set up in the schools where the subjects were available for testing. The instructions were provided to small groups of four or less, but subjects were run in private. The mechanics of the test were explained in the instructions, along with advice that the test was being used for research purposes only and that it was not a timed test but the subject should work quickly without rushing. After the subject's task had been described, they were shown a two-item test not using the figures in the record test to demonstrate how the test would be run and to acquaint the subject with nonsense figures. The subjects were then run individually. Once the first stimulus was presented, the test ran continuously with no apparent break until the 60th frame had been processed.

Table 2

Subtests of the ASBAV Form 5

Name of Test	Number of Items	Subtest Reliabilities*
(GI) General Information	15	.67
(NO) Numerical Operations	50	.85
(AD) Attention to Detail	30	.82
(WK) Word Knowledge	30	.91
(AR) Arithmetic Reasoning	20	.82
(SP) Space Perception	20	.82
(MK) Mathematical Knowledge	20	.88
(EI) Electronic Information	30	.87
(MC) Mechanical Comprehension	20	.81
(GS) General Science	20	.77
(SI) Shop Information	20	.83
(AI) Automotive Information	20	.84

*The data are from Jensen, et al., (1977). The reliabilities were derived using Kuder-Richardson Formula 20 with the exception of Numerical Operations and Attention to Detail, which were obtained by test-retest methods using ASVAB Form 6.

RESULTS

The total exposure time of the stimuli ranged from 35.5 to 161.1 sec. with a mean exposure time of 79.1 sec. Incorporating the 1.4-sec. cycle time between stimuli, the individual administration of the test required an average of 2.7 min. Since all subjects were administered 60 stimulus pairs, those with the shortest exposure times were averaging a little over .5 sec. per frame. Speed on the test could be a characteristic of quick learning or a rapid response set. The latter might be the result of negative motivational factors induced by telling the subjects that the test was being given for strictly research purposes. Questions about the role of rate of responding carried considerable concern, since the scoring of the test was in terms of the number of correct responses per unit of viewing time. This was called the Information Processing Rate (IPR) since each stimulus pair carried one bit of information. The correlation between the number correct and viewing time was -.73, which was significant at the .01 level. This indicated that the indivudals who learned more required less time. Accordingly, it was concluded that subjects were motivated to perform well and that quick responding was, as originally hypothesized, an indication of rapid learning.

The means and standard deviation on all subtests of the ASVAB, the AFQT composite, and the IPR are shown in Table 3 by sex, ethnic group, and the entire sample. The IPR was multiplied by 1,000 for convenience in displaying the ratio.

At the .05 significance level, there were no male-female differences in IPR scores for the total sample or the subsamples. There were significant differences between all whites and nonwhites (_t_ = 2.20) and between white and nonwhite females (_t_ = 2.30). The difference of 72.42 in the mean scores of white and nonwhite males did not reach statistical significance. Thus, it appears that there are white-nonwhite differences in IPR performance, and that this difference was due primarily to differences between females of the two groups.

On the AFQT, there was a significant difference in mean scores between males and females at the .05 level for only the white subjects (_t_ = 2.26). No differences were found between all males and females and between nonwhite males and females. There were significant white-nonwhite differences in mean AFQT scores for all categories of subjects. The white-nonwhite difference for all subjects was significant at the .01 level (_t_ = 3.00). The differences between white and nonwhite males (_t_ = 2.44) and between white and nonwhite females (_t_ = 2.10) were signigicant at the .05 level. To summarize, there are consistent differences between all white and nonwhite groupings on the AFQT dimension. The only sex-related difference occurs between male and female whites.

Because of the differences in the sizes of the subsamples, the _t_-test was used to assess the differences for each contrast rather than an analysis of variance incorporating all of the variables simultaneously. In the significant differences that were found, the higher mean was always for whites or males.

The correlation of the IPR score with the ASVAB tests and the AFQT composite are shown in Table 4 for the total sample and by sex and ethnic groups. The most noteworthy correlations in Table 4 are those between IPR and General Information (GI) for the total sample and for nonwhites at a significance level of .01 and for females at a significance level of .05. The correlation of IPR with Mechanical Comprehension (MC) followed a similar pattern, except that the correlation was not as high, and for females, the correlation of .31 was significant at only the .06 level. There was also a low, but significant, correlation of IPR with AFQT for the total sample and females. There is a complete absence of correlation between IPR and the psychometric test variables for whites and males. In the case of the former, General Information (GI) and Automotive Information (AI) are the highest correlations, while General Information and Mechanical Comprehension (MC) are the highest for the males. Thus, the nonwhites and females appear to be the prime contributors to any obtained relationship between the IPR scores and the psychometric test variables.

In order to obtain an indication of the relationship to IPR of all of the variables in the study considered simultaneously, the IPR scores were regressed in a stepwise manner on the study variables using the SPSS program

TABLE 3

MEAN TEST SCORES BY RACE AND SEX - ALL SCHOOLS

	WHITE			NONWHITE			TOTAL		
	Male	Female	Total	Male	Female	Total	Male	Female	Total
N	17	26	43	11	11	22	28	37	65
GI	10.29 1.69	7.85 1.83	8.81 2.13	10.09 2.95	6.46 2.12	8.27 3.12	10.21 2.22	7.43 1.99	8.63 2.50
WK	21.88 5.08	17.89 5.60	19.46 5.69	17.36 7.49	13.46 6.79	15.41 7.26	20.11 6.41	16.57 6.23	18.09 6.50
MK	14.47 4.19	13.15 4.32	13.67 4.27	10.64 4.43	11.73 5.41	11.1 4.86	12.96 4.62	12.73 4.64	12.83 4.59
GS	12.06 4.01	8.92 3.03	10.16 3.74	8.91 2.81	6.73 3.04	7.82 3.07	10.82 3.86	8.27 3.16	9.37 3.68
NO	36.53 7.98	36.50 8.05	36.51 7.92	31.91 9.75	36.09 11.40	34.00 10.57	34.71 8.84	36.38 9.00	35.66 8.90
AR	13.47 4.24	11.96 3.18	12.56 3.67	10.64 4.23	10.00 3.19	10.32 3.67	12.36 4.39	11.38 3.27	11.80 3.79
EI	17.24 5.87	12.31 4.10	14.26 5.39	14.64 4.52	12.82 2.82	13.73 3.80	16.21 5.45	12.46 3.73	14.08 4.88
SI	12.88 3.77	8.92 2.56	10.49 3.63	11.64 3.78	6.91 2.17	9.27 3.86	12.39 3.76	8.32 2.59	10.08 3.72
AD	13.71 3.87	14.73 3.09	14.33 3.41	14.64 3.33	13.09 4.78	13.86 4.10	14.07 3.63	14.24 3.69	14.17 3.63
SP	12.41 5.41	10.58 3.69	11.30 4.48	8.46 3.39	9.00 4.38	8.73 3.83	10.86 5.05	10.11 3.91	10.43 4.42
MC	11.94 3.60	8.35 3.05	9.77 3.69	9.82 2.68	5.82 1.17	7.82 2.87	11.11 3.38	7.60 2.86	9.11 3.54
AI	9.35 4.89	7.15 3.03	8.02 3.97	8.91 2.91	5.82 2.27	7.36 3.00	9.18 4.16	6.76 2.86	7.80 3.66
AFQT	47.77 11.36	40.42 9.73	43.33 10.89	36.45 12.45	32.46 12.36	34.46 12.28	43.32 12.87	38.05 11.03	40.32 12.05
IPR	659.06 248.80	648.00 265.64	652.37 256.15	586.64 176.89	450.00 152.52	518.32 175.69	630.60 222.64	589.14 252.75	607.00 239.32

Note: Top number is test mean. Bottom number is test standard deviation. The table is from Sherman (1979).

TABLE 4

Correlations of IPR with ASVAB Tests and AFQT Composite

	ASVAB Subtests**												
Group	GI	WK	MK	GS	NO	AR	EI	SI	AD	SP	MC	AI	AFQT
Combined	.34*	.25	.20	.21	.00	.18	.03	.23	.08	.19	.28	.22	.26
White	.26	.14	.14	.06	-.03	.09	.03	.19	.00	.14	.16	.27	.16
Nonwhite	.50*	.33	.17	.44	-.08	.18	-.05	.26	.26	.08	.48	-.03	.27
Male	.34	.00	.08	.20	.05	.11	.00	.14	.18	.14	.24	.10	.10
Female	.36	.40	.28	.19	-.02	.23	-.01	.32	.03	.22	.31	.31	.37

*Significant at $p \leq .01$.

_Significant at $p \leq .05$.

(Data are from Sherman, 1979)

**See Table 3 for full test titles.

(Nie, et al., 1975). The independent variables included the ASVAB tests, the AFQT composite, two dummy variables for the three high schools, a dummy variable for ethnic group, and a dummy variable for sex. Interactive variables were created by multiplying the General Information and AFQT scores by each of the dummy variables. The stepwise procedure was stopped when the adjusted r^2 did not improve and the significance of the overall F ratio for regression failed to improve. The fitted equation is shown in Table 5.

Table 5
Stepwise Regression of IPR on the Study Variables

Variables in Equation*	B	Beta	Std Error B	F	Sig
GI	32.22	.34	11.83	7.42	.01
AFD3	3.48	.33	1.28	7.33	.01
G1D1	-12.49	-.21	6.91	3.27	N.S.
EI	- 8.67	-.18	6.16	1.98	N.S.
constant 384.96					

*See test for identification of the variables.

With 4, 60 d.f., the obtained F ratio of 4.7 for regression was significant at the .005 level. It should be noted, however, that other interpretations of the r^2 in the stepwise regression might not consider the obtained r^2 to be statistical significant (Wilkinson, 1979).

The variables in the equation included General Information (GI); an interactive variable, AFQT times D3, the race dummy (1 = white, 0 = nonwhite); GI times a school dummy; and EI (Electronics Information). Only the first two contributed to the equation at a statistically significant level. Thus, for all subjects, GI was the best predictor of IPR and for whites, the AFQT was also a significant predictor. The latter would seem to incorporate the fact that whites scored higher than nonwhites on both the IPR and the AFQT. The latter was the best variable to scale the difference between whites and nonwhites on the IPR.

DISCUSSION

Comparison with Previous Findings

One of the objectives of the study was to compare its findings with the results of the original study using the discrimination learning test (Arima, 1979). The IPR in the previous study for the self-paced condition was 216.5. The IPR in the present study was 607.0. The possible sources of the difference are too many to make reliable comparisons. However, the two items that stand out are the automation of the present version vs. the manual advance of the earlier test and the correlation method (contingent reinforcement) used in the present study. In the present study, the subject had to press the correct alternative to advance the system, whereas the subject in the former study was merely informed by a light when he or she made the correct choice by depressing the appropriate response buttons.

There were significant white-nonwhite differences in the machine-paced condition of the earlier study that apparently disappeared in the self-paced mode. There are still significant white-nonwhite differences, but the primary contribution to this difference comes from the female subjects where there was a 200-point difference favoring the white females. Since there were no sex differences among the white subjects, and there was a 136-point difference between male and female nonwhite subjects (Table 3), it appears that the nonwhite females were a particularly low-performing sample. There were no females in the previous study and no significant difference between white and nonwhite male subjects in the present study. Accordingly, there is some justification for concluding that there are no reliable differences between white and nonwhite male subjects. More data would be required to make a similar statement for the female subjects.

In the earlier study, there was a statistically significant correlation between IPR and the AFQT for the total sample and the white subsample. The correlation was not significant for the nonwhites. In this study, there is still a significant relationship between the IPR and AFQT for the total sample, but the significant subsample correlation now occurs in the female subsample. Nevertheless, in view of the repetition of the significant correlation for the larger (total) sample and the regression equation in which, as formerly, the AFQT plays a significant role for only the white subjects, it is concluded that there is modest, but reliable, relationship between the learning performance and the AFQT score. This relationship is further explored below.

Relationships between Learning Performance and ASVAB Test Scores

The relationship between learning performance and the psychometrically derived ASVAB test scores is of particular interest to this study. There is no doubt that a close relationship exists between the IPR scores and GI (General Information). This is evident in the degree and pattern of correlations seen in Table 4 and in the regression equation in Table 5. As previously stated, GI is a test instigated by the Army to provide a "bottom" to the ASVAB. The Army needed a test to differentiate the potential usefulness of individuals who score low on the basic tests used for screening enlistees. In the present study, the highest correlations between IPR and GI scores occurred for those subsamples scoring lower on the AFQT--nonwhites and females. For subjects scoring higher on the AFQT, it may be that ceiling effects in both variables attenuated the calculated relationship (correlation) between them.

To explore the IPR-GI relationship further, the nature of GI, itself, should be examined. First, Table 3 shows that there is no white-nonwhite difference in the GI scores. This comparison holds up very well when the comparison is made between white and nonwhite males and white and nonwhite females. There appears to be a large difference in GI between males and females, just as there is in the AFQT scores. It is remarkable--considering that the other tests of the ASVAB are longer, more reliable, and generally recognized to be the "heavy-weights" in evaluating individuals--that the 15-item GI test should stand out as the best predictor of learning performance. The relationship of GI to the other ASVAB tests is shown in Table 6. The table reveals that GI is significantly correlated with every other subtest in the battery, and it is also one of three tests that identify the first factor (Verbal) extracted from the test correlations (U.S. Enlisted Processing Command, undated). The factor was identified as the ability to tie words and information together. The foregoing would seem to justify the contention that GI is a measure of a strong general factor that pervades and dominates the ASVAB tests and especially the composites (Cronbach, 1979).

Table 6

CORRELATION BETWEEN GENERAL INFORMATION AND OTHER ASVAB SUBTESTS*

NO	AD	WK	AR	SP	MK	EI	MC	GS	SI	AI
.44	27	61	52	34	52	61	57	59	61	57
28	14	52	47	34	43	53	51	49	50	47

*Based on Service standardization sample (upper row) and sample of 2,052 students in the 10th, 11th and 12th grades (bottom row).

As for the IPR score, it has only been identified as a rote learning score. It is not a perceptual or speed test, as evidenced by the zero or near-zero correlations between it and the Attention to Detail and Numerical Operations subtests. It is also not dependent on spatial perception as demonstrated by a relatively low correlation with the SP test in Table 4. It is correlated, for the general sample, with Word Knowledge, Mechanical Comprehension, and the AFQT. On the basis of the differential test results, the IPR score is apparently the result of coding (labeling), organizing, and storing in short-term memory for immediate retrieval discriminating, information about the nonsense form, stimulus pairs. Jensen (1979) states that this sort of a task makes moderate demands on the concept he calls g, a general measure of mental ability or intelligence.

From the preceding analysis of the characteristics of the GI test and the IPR measure, it is hypothesized that they are both measuring a general capacity for processing and using information and a general characteristic of alertness and responsivity to the environment. One would conjecture that either measure would be related to the latency of the alerting response as measured in recent studies using averaged brain potential responses to a light stimulus. These concepts require experimental verification, of course.

As a performance measure, the learning task could be improved and made more discriminating of individual differences if an individually determined stopping criterion were used. For example, 12 successive, correct responses might be such a criterion. The intent was to investigate this possibility, but the instrumentation proved to be uncooperative. A fixed number of trials, as well as paced presentations, penalizes the rapid learner. The information processing rate should be calculated for the learning period and not attenuated by the time required for reflexive responding once the material has been learned.

Implications for Personnel Selection

If the IPR were scored with an individual stopping criterion in order to increase the variance in performance among individuals, it would seem to be an effective and efficient measure of the general intelligence of a person that is reasonably culture free. While it apparently measures the same area as a general factor that dominates the ASVAB, it provides the opportunity for those with poorer language skills to show their capabilities in the areas of the highly language-dominated tests of the ASVAB. With the advent of computerized testing, this and similar performance tests should be simple and efficient to administer and could provide a greater pool of individuals for selection. Moreover, there has been little validation of the selection instruments with performance in the Armed Services because positive correlations are typically not found. It could be that simple performance tests used as selectors might provide the dimensions to better the validation of selection tests. In the area of truth in testing, the performance tests would have a great advantage since the correct answers could be tailored for each subject at the time of testing if the tasks were designed to permit this option. For example, in the present discrimination learning test, the correct number of each pair could be randomly determined immediately prior to testing.

If the ASVAB composites are so dominated by the general factor to make them essentially useless for counseling as asserted by Cronbach (1979), the same could be said for their use in placement, as employed by the Armed Services. Reliable differences must exist between the composites to make either function possible. Unfortunately, the correlations among the key technical Navy composites range from .88 to .91. Swanson (1978) provides validation data for end-of-course grades or time-to-completion of self-paced courses for 19 schools using the General Technical composite and 8 schools using the Mechanical composite. In almost all of the cases, the correlations are higher for the Electronics composite. The Electronics composite holds up well as the selector with the highest correlation for the 9 schools using it as a selector. Judging from these limited examples, it would be more efficient just to use the Electronics composite as the selector for all of the schools shown in Swanson's study. This study has served to reinforce the notion of a general factor dominating the ASVAB tests by calling attention to the pervasive relationship of General Information to all of the tests and the fact that the General Information test best predicts scores on a discrimination-learning, performance test.

Finally, attention should be called to the case of the females in this study. They are typical of standardization populations in general for the ASVAB (Jensen, 1977) in that their AFQT scores are one-half a standard deviation lower than the males, and they do poorly in the trade tests. If the standardization norms are strictly applied, the females are very adversely affected in selection for service or the more desirable technical courses. They maintain equity only in the areas of Attention to Detail and Numerical Operations that are the key elements of the Clerical composite. It should be noted again that the mean IPR scores of the white males and females were identical, indicating that they were comparable in general cognitive ability.

REFERENCES

Arima, J.K. A culture-free performance test of learning aptitude (NPS-54-78-2). JSAS Catalog of Selected Documents in Psychology, 1978, 8, 93. (MS 1775)

Bitterman, M.E. The evolution of intelligence. Scientific American, January 1965, 92-100.

Bitterman, M.E. The comparative analysis of learning. Science, 1975, 188, 699-710.

Carroll, J.B., and S.E. Maxwell. Individual differences in cognitive abilities. Annual Review of Psychology, 1979, 30, 603-640.

Cronbach, L.J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.

Cronbach, L.J. The Armed Services Vocational Aptitude Battery--A test battery in transition. Personnel and Guidance Journal, 1979, 232-237.

Estes, W.K. Learning theory and intelligence. American Psychologist, 1974, 29, 740-749.

Jensen, A.R. The nature of intelligence and its relation to learning. Journal of Research and Development in Education, 1979, 12, 79-95.

Jensen, H.E., I.H. Massey, and L.D. Valentine, Jr. Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7) (Technical Research Note 77-3). Fort Sheridan, IL: U.S. Army Enlisted Processing Command, 1977.

Kettner, N. Armed Services Vocational Aptitude Battery (ASVAB Form 5): Comparison with GATB and DAT tests (Technical Research Report 77-1). Fort Sheridan, IL: U.S. Army Enlisted Processing Command, 1977.

Nie, N.H., C.H. Hull, J.G. Jenkins, K. Steinbrenner & D.H. Bent. Statistical package for the social sciences. (2nd ed.) New York: McGraw-Hill, 1975.

Sherman, V.A., Jr. Culture-free testing for selection of recruits. Unpublished master's thesis, Naval Postgraduate School, 1979.

Smith, R.J. "Truth-in-testing" attracts diverse support. Science, 1979, 205, 1110-1114.

Sternberg, R.J. The nature of mental abilities. American Psychologist, 1979, 34, 214-230.

Swanson, L. Armed Services Vocational Aptitude Battery Forms 6 and 7: validation against school performance--interim report (NPRDC TR 78-24). San Diego, Calif: Navy Personnel Research and Development Center, 1978.

U.S. Military Enlisted Processing Command. ASVAB counselor's guide. Fort Sheridan, IL: Author, undated.

Wilkinson, L. Tests of significance in stepwise regression. Psychological Bulletin, 1979, 86, 168-173.

Young, P.A. A culture-free performance test of general learning ability. Unpublished master's thesis, Naval Postgraduate School, Monterey, CA, December 1975.

STUDYING INDIVIDUAL DIFFERENCES-JOB PERFORMANCE LINKS IN MILITARY RECRUITER PERFORMANCE: A CONSTRUCT VALIDITY APPROACH

Walter C. Borman
Rodney L. Rosse
Personnel Decisions Research Institute

Norman M. Abrahams
Navy Personnel Research and Development Center

and

Jody L. Toquam
Personnel Decisions Research Institute

INTRODUCTION

This paper describes the third phase of a research program to develop paper-and-pencil predictors of Navy recruiter performance. Previous research within this program has resulted in behavior-based performance rating scales tailor-made for evaluating Navy recruiter effectiveness (Borman, Hough & Dunnette, 1976) and a paper-and-pencil battery of predictor scales with demonstrated concurrent validity (Borman, Toquam & Rosse, 1977).

The purpose of this program phase has been to identify possible underlying personality and vocational interest constructs that might be associated with effective recruiter performance and then to add items that ıp these same valid constructs in an effort to better understand the co.ıstructs and to increase the validity of the resulting battery. This paper describes a) steps taken to identify potentially valid constructs in the personality and vocational interest domains, b) the addition of items leading to a test of our understanding of these constructs and hopefully toward improvements in the measurement of them, and c) results of a study designed to evaluate the validity of the constructs as indicators of recruiter performance.

METHOD

Development of "Valid Item Pools"

From the trial inventory battery developed during previous research (Borman et al., 1977), we first selected personality and vocational interest items that proved to be valid in a concurrent validity study according to criteria explained below (N = 267; Borman et al., 1977). Separate item pools were formed for each of the four performance categories in that validity study (Selling Skills, Human Relations Skills, Organizing

Skills, Overall Performance), and personality and vocational interest items were also kept separate, this procedure resulting in eight different item pools.[1]

The criterion for inclusion of personality items in a pool was a correlation of $|.10|$ or greater with the target performance category in the $N = 267$ sample. Of the 310 personality items in the trial inventory battery, the number selected was, respectively, 55, 85, 95, and 80 for the four performance categories named above. For the vocational interest items, selection for a pool was predicated on the following decision rule: $r > |.12|$ in $N = 267$ sample and $r > .00$ in the same direction in $N = 62$ pilot test sample (from previous study); or $r > |.24|$ in $N = 62$ sample and $r > \pm.08$ in the same direction for the $N = 267$ sample. This screening procedure resulted in, respectively, 39, 62, 48 and 53 items for the four performance categories from the total of 325 vocational interest items included in the trial inventory battery.

Factor Analyses to Identify Valid Constructs

In the $N = 267$ sample, responses to the personality items were intercorrelated within item pool, and the resulting correlation matrices factor analyzed. Two to ten factors were extracted and each solution was rotated to the varimax criterion. The same procedures were employed with the vocational interest item pools. Thus, eight different sets of factor solutions were generated, and the most psychologically meaningful solution was selected for each item type (personality or vocational interest) and each performance category. Names and definitions of the factors are presented in Tables 1 and 2.

Table 1

Factor Analysis Results for Personality Items:
The "Valid Constructs"

Selling Skills

I Good Impression
II Impulsive; carefree *vs.* order; planning ahead; systematic; levelheaded
III Enjoyment of being center of attention, leading, showing off, and speaking before a group
IV Working hard and with confidence; being happy *vs.* being unhappy; giving up easily; disgruntled about life

[1] Items were not constrained to appear in only a single pool. In fact, some items appeared in all four item pools.

Table I (continued)

Human Relations Skills

I Preference for working with and being with people
II Spontaneity; impulsivity; "fast and careless"; rebellious; tendency to have bad moods
III Unhappy; lack of confidence; disgruntled about life[a]
IV Ambitious; working hard; pushing self

Organizing Skills

I Order; planning ahead; well organized _vs_. impulsive; acting without thinking; "fast and careless"
II Leading and influencing others; giving orders; demanding of self; ambitious; dominant
III Unhappy; discouraged; doing little in life; giving up hope; feeling useless[a]
IV "Bad actor"; was unruly and rebellious in school; unsocialized[a]

Overall Performance

I Doing more than expected _vs_. giving up; working just hard enough
II Impulsive; "fast and careless" _vs_. order; methodological; planning ahead
III Leading and influencing others; dominant; strong personality
IV Good impression _vs_. admitting occasional meanness, grouchiness, disgust with self, discouragement uselessness, bad mood
V People oriented; liking to be around others and close to others; open to other people

[a] These three personality constructs related _negatively_ to their target performance criteria.

Table 2

Factor Analysis Results for Vocational Interest Items:
The "Valid Constructs"

Selling Skills

I Interests in extroverted, dominant, leadership activities and occupations
II Interest in occupations involving attention to detail[a]
III Interests in law and politics
IV Interest in sports and competitive activities

Table 2 (continued)

Human Relations Skills

I Interests in dominant, extroverted, social activities
II Interests in teaching and counseling
III Interests in "feminine" occupations and activities
IV Interests in newspaper reporting and foreign service
V Interest in sports and competitive activities
VI Interest in religion and in being around the sickly

Organizing Skills

I Interests in politics and high level management jobs
II Interests in bookkeeping, statistical, and detail work
III Interests in "feminine" occupations and activities[a]
IV Interests in leadership and responsibility

Overall Performance

I Interests in law and politics, and management occupations and activities
II Interest in activities and occupations that require extroversion, dominance, responsibility, and leadership
III Interest in sports and competitive activities
IV Interests in teaching and counseling
V Interest in "feminine" occupations[a]

[a] These three vocational interest constructs related *negatively* to their target performance criteria.

Generating New Items Targeted Toward Valid Constructs

To generate new personality items, we first reviewed several personality inventories for scales related conceptually to one or more of the 17 constructs identified in the factor analyses. After selecting personality scales that seemed to reflect constructs discovered in the factor analyses, we searched for items within these scales that appeared, in particular, to tap those same constructs. For example, for the construct, Leading and Influencing Others, we selected the item, "I try to control others rather than permit them to control me" from the Dominance scale of the Personality Research Form. In all, 83 items were selected and targeted toward constructs from the factor analyses. In addition, we *wrote* 26 personality items, again, targeted toward the valid constructs. The latter items were of the same general type as the item offered as an example above.

In the vocational interest domain, 49 items were written to tap the constructs derived from the factor analyses of valid vocational interest items. For example, the item, "College football coach" (to which the respondent answers like, indifferent, or dislike) was written to measure the constructs related to sports interests, and the item, "Keeping track of statistics for baseball, football, etc.", was written to tap the Interests in Detail Work construct.

Investigating the Relationships Between the Individual Differences Constructs and Recruiter Performance

Personality and vocational interest items found to be valid in our previous study (Borman et al., 1977), along with all of the new items selected or written to measure the valid constructs, were administered to a nationwide sample of 194 Navy recruiters working in seven different cities or suburban areas.

Also administered were the 17 rating scales used in that study. Recruiters evaluated themselves and other recruiters with whom they worked closely, and supervisory personnel evaluated the recruiters who worked for them. Thus, as in our previous study, performance ratings were gathered from three different sources, these ratings yielding the criterion performance scores against which we evaluated the validity of the personality and vocational interest constructs. Then, correlational analyses, described in the Results section below, were conducted a) to assess how precisely the new items measured their target constructs and therefore how clearly we understood these constructs; and b) to evaluate the validity of the constructs as indicators of Navy recruiter effectiveness.

RESULTS

This section first describes development of criterion performance scores for recruiters in the present sample (N = 194) and then outlines results of the validity analyses.

Criterion Development Analyses

Interrater reliabilities were examined for the following rating sources: peer, supervisor-peer, supervisor-self, peer-self, and supervisor-peer-self. Also, for each of these rating sources, we investigated the underlying dimensionality of the ratings by intercorrelating the dimensions and factor analyzing each of the resulting correlation matrices. Based on the magnitude of the reliabilities and the meaningfulness of the factor solutions, the peer and supervisory ratings, pooled together, were selected for further criterion development analyses.

The means for the 16 dimensions and the Overall Performance dimension range from 6.47 to 7.48 with a median of 7.03 (on a 1-10 scale), suggesting that leniency error is not a serious problem here. The standard deviations range from 1.05 to 1.56 with a median of 1.32, suggesting that the range of the ratings is not severely restricted. And finally, most reliabilities appear acceptable (.34 to .77, median r=.54) and at a level comparable to our previous recruiter study (Borman et al., 1977).

Factor analyses of the performance ratings indicated that the three factor solution is the most psychologically meaningful. This solution is also strikingly similar to solutions generated previously (Borman et al., 1977), lending still more support to the stability of this dimensional structure for describing Navy recruiter performance. The factors are: Selling Skills, Human Relation Skills, and Organizing Skills.

Factor scores were computed for each recruiter in the sample, and the interrater reliabilities were, respectively, .62, .48, and .65, sufficiently high to allow the factor scores to represent individual recruiters' effectiveness in three different aspects of Navy recruiting. Therefore, factor scores on the three composite categories, reflecting three conceptually meaningful aspects of recruiter performance, and the highly reliable overall performance rating, providing a summary effectiveness measure, were used as criteria to evaluate relationships between the various personality/vocational interest constructs and recruiter performance.

Relationships Between Individual Differences Constructs and Performance: The Personality Domain

Appearing in Table 3 are validities of composites consisting of old items, new items, and old plus new items together, these composites[2] measuring the constructs identified in our factor analytic research described above. Also presented in this table are convergent validity indices, correlations between the old and new item composites targeted to measure the same constructs.

Focusing first on the question of our understanding of these various personality constructs, notice that the new items do appear to be tapping substantially the same constructs as the old items. The median correlation between old and new item composites measuring the 17 constructs is .56 ($p < .001$), indicating reasonably high convergent validity for these pairs of composites. Further, discriminant validity of the new item composites is good; convergent validity correlations are greater than correlations between each new item composite and old item composites measuring

[2] Each "old item" composite corresponding to a factor (construct) was formed by unit weighting responses to all items loading sufficiently highly on that factor (and not highly on any other factor) in the N = 267 sample. In other words, an old item composite for a factor (construct) consisted of the marker items for that factor, these items unit weighted to form the composite. Each new item composite (one for each construct) was developed by simply unit weighting responses to the new items targeted toward the construct.

Table 3

Relationships Between Old and New Personality Item Composites and Validities of These Composites Against Recruiter Performance

					Validities Against Performance			
Performance Category	Construct[a]	No. of Old Items	No. of New Items	Convergent Validities: Correlations Between Old and New Item Composites	Old Item Composites in Previous Sample	Old Item Composites in Present Sample	New Item Composites in Present Sample	Old Plus New Item Composite in Present Sample
Selling	Good Impression	4	4	32[c,d]	18	20	05	18
	Impulsive	15	32	72	23	01	03	03
	Leading and "showing off"	18	31	59	29	26	22	27
	Working hard	15	21	53	30	17	05	11
Human Relations	People oriented	13	34	56	23	12	08	11
	Spontaneity	7	30	14	37	08	23	22
	Unhappy	7	6	40	-18	-14	-16	-17
	Working hard	10	15	46	30	24	26	29
Organizing	Order	25	37	77	31	12	22	17
	Leading and influencing	16	32	64	26	10	12	12
	Unhappy	13	10	39	-28	01	04	00
	"Bad actor"	6	10	21	-14	-05	-04	06
Overall Performance	Working hard	8	12	51	28	05	16	13
	Impulsive	10	27	61	15	05	07	08
	Leading and influencing	13	37	64	28	26	31	33
	Good impression	9	11	62	25	12	-02	08
	People oriented	11	35	57	26	07	-04	00
All Performance Category Constructs together[b]	Selling	85	88	73	47	25	19	23
	Human Relations	55	85	57	44	18	25	24
	Organizing	95	89	69	45	09	22	15
	Overall Performance	80	122	75	46	21	21	22

[a] See Table 1 for more complete definitions of these constructs.

[b] For each performance category, a single composite was formed by unit weighting all items contained in the four or five composites targeted toward that criterion.

[c] $P_{.05} = .14$, $P_{.01} = .19$ for correlations in this table.

[d] Decimal points are omitted for all correlations in this table.

different constructs for all but two of 56 such comparisons.[3] These results mean that in most cases we successfully conceptualized the constructs and were able to write new items focused directly toward those constructs.

Next, note the validities for the old item composites in the previous sample. These should, of course, be high because the items contained in those composites were selected in part according to their validities in that sample. Validity coefficients of old item composites in the present sample (N = 194) can then be viewed as cross-validity estimates for those composites. Unfortunately, many of these cross-validities are much reduced in magnitude, but several suggest reasonably high and consistent relationships between the constructs and performance.

Regarding the validity of the new items, Table 3 data indicates that all but two of the 17 validities for the new item composites are in the proper direction, and seven of these 17 validity coefficients are significantly different from zero at the .05 level or greater (and in the predicted direction). Therefore, these new items not only appear to measure substantially the same constructs as was intended, but also they contribute, in the main, at least modest validity against their target performance criteria. In fact, for a little over half the constructs (9 of 17) of the new item composites provide validities higher than those provided by the old item composites. Also, for nine of the 17 constructs, composites consisting of the old plus the new items together show higher validities than do the composites containing old items alone. In other words, including the new items in composites enhances the validity of those composites for slightly more than half the constructs.

Relationships Between Individual Differences Constructs and Performance: The Vocational Interest Domain

Analyses for vocational interest composites were conducted in the same manner as those just described for the personality composites, and Table 4 presents the results. Again, the new items written to measure the target constructs in fact appear to be successfully measuring those constructs. The median correlation between the new item composites and the old item composites (the convergent validities) is .67 ($p < .001$). Also, discriminant validity is excellent; in all 66 comparisons between convergent validities and correlations between new and old item composites not intended to measure the same construct, the magnitude of the convergent validities is greater. Thus, as was the case in the personality domain,

[3] To help explain the discriminant validity analysis: For example, the .32 convergent validity coefficient for the Good Impression construct is greater than the correlations between the new item composite for Good Impression and the old item composites representing a) Impulsiveness, b) Leading and "Showing Off", and c) Working Hard, and this same pattern of discriminant validity obtains for almost all of the other new item composites.

Table 4

Relationships Between Old and New Vocational Interest Item Composites and Validities of These Composites Against Recruiter Performance

					Validities Against Performance			
Performance Category	Construct[a]	No. of Old Items	No. of New Items	Convergent Validities: Correlations Between Old and New Item Composites	Old Item Composites in Previous Sample	Old Item Composites in Present Sample	New Item Composites in Present Sample	Old Plus New Item Composite in Present Sample
Selling	Extroverted interests	9	7	69[c,d]	25	20	20	22
	Interests in detail work	5	2	65	-18	00	-03	-01
	Law and political interests	11	2	75	21	13	15	14
	Sports interests	6	6	51	23	-03	03	00
Human Relations	Extroverted interests	16	4	56	23	23	19	24
	Interests in teaching	4	1	78	17	13	07	11
	"Feminine" interests	8	5	53	24	05	07	07
	Interests in newspaper work	6	2	58	18	16	06	14
	Sports interests	3	6	62	17	20	20	22
	Religious interests	5	4	71	22	09	13	12
Organizing	Interests in politics	16	4	76	23	16	11	15
	Interests in detail work	10	3	50	24	03	-01	02
	"Feminine" interests	4	2	52	-19	08	-02	05
	Leadership interests	7	4	58	29	01	-04	-01
Overall Performance	Law and political interests	11	4	71	21	10	21	14
	Extroverted interests	14	7	69	28	23	22	24
	Sports interests	6	6	72	21	13	05	10
	Interests in teaching	5	2	73	19	11	00	08
	Feminine interests	3	0	--	-14	-03	--	-03
All Performance Category Constructs together[b]	Selling	39	17	62	35	19	20	21
	Human Relations	62	22	69	32	20	21	22
	Organizing	48	13	63	37	14	04	12
	Overall Performance	53	19	73	23	21	19	22

[a] See Table 1 for more complete definitions of these constructs.

[b] For <u>each</u> <u>performance</u> <u>category</u>, a single composite was formed by unit weighting all items contained in the four, five, or six composites targeted toward that criterion.

[c] $P_{.05} = .14$, $P_{.01} = .19$ for correlations in this table.

[d] Decimal points are omitted for all correlations in this table.

the levels of convergent and discriminant validity noted here indicate accurate conceptualization and relatively precise measurement of several vocational interest constructs.

Finally, the pattern of validities appearing in Table 4 is very similar to the pattern noted with the personality items. Old item composites, of course, related well to performance in the original (development) sample, and in general, these relationships are lower in the present (N = 194) sample. About half the validities for the new item composites are higher than the validity coefficients provided by old item composites (seven of 16, two ties). And, for 11 of the 18 constructs considered here, validities of composites consisting of old and new items pooled together are higher (in the intended direction) than the validity coefficients obtained when the old items alone comprise the composites.

DISCUSSION AND CONCLUSIONS

The primary purpose of this research has been to study the individual differences constructs related to Navy recruiter performance. The procedures presented here succeeded in identifying personality and vocational interest constructs potentially related to one or more aspects of recruiter effectiveness, and attempts to develop additional measures of these constructs were reasonably successful. These results were interpreted to mean that our level of understanding of many of these constructs is high, because we have, essentially, developed parallel forms of the construct measures.

Further, relationships between composites of the old plus the new personality/vocational interest items and performance on target effectiveness criteria were reasonably high for several constructs, providing confirmation of stable links between individuals' scores on these constructs and recruiter performance. These stable and comparatively well understood personality and vocational interest constructs were then presented as individual differences highly likely to differentiate between effective and ineffective performance in Navy recruiting.

We conclude that the general process presented here of attempting to discover, understand, and then confirm individual differences constructs important for job performance has considerable merit. It provides one way to gain understanding of individual differences--job performance linkages, and this understanding should contribute to more precise prediction of Navy recruiter effectiveness.

REFERENCES

Borman, W. C., Hough, L. M., & Dunnette, M. D. Development of behaviorally based rating scales for evaluating the performance of U. S. Navy recruiters. (NPRDC TR 76-31) San Diego, Calif.: Navy Personnel Research and Development Center, February, 1976.

Borman, W. C., Toquam, J. L., & Rosse, R. L. Development and validation of an inventory battery to predict Navy and Marine Corps Recruiter Performance. Final report for Contract Number N00123-76-C-1284, submitted to Navy Personnel Research and Development Center. Minneapolis, Minn.: Personnel Decisions Research Institute, January 1978.

INITIAL DEVELOPMENT OF OPERATIONAL COMPOSITES FOR THE VOCATIONAL INTEREST-CAREER EXAMINATION

Thomas W. Watson, William E. Alley, and Mary E. Southern

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

INTRODUCTION

For many employers, especially firms requiring a large number of employees for diverse jobs, initial job classification is a difficult process, especially when an optimum match between a job incumbent and a career field is an important goal. When job applicants are relatively young, vocationally inexperienced, and untrained, the process of initial job placement becomes particularly difficult. The Air Force recruits thousands of young men and women annually and places them into hundreds of career fields. Historically, the Air Force has used two primary criteria for making initial classification decisions: the aptitude of the individual, and Air Force requirements at the time of entry.

The Air Force has developed and implemented a computerized job reservation system which greatly enhances the ability of the Air Force to consider information other than aptitude when job placement decisions are being made (Ward, Haney, Hendrix, & Pina, 1978). Vocational interest data represent one such additional type of information. It is assumed that use of vocational interest information as an adjunct to aptitude data will provide positive individual and organizational consequences by increasing job satisfaction, productivity, and retention.

This paper describes the initial development of an operational scoring procedure for the Vocational Interest-Career Examination (VOICE), a general purpose vocational interest inventory intended primarily for use with Air Force enlisted personnel. The inventory was initially produced under contract by the Educational Testing Service (Echternacht, Reilly, & McCaffrey, 1973) and later further refined and validated on Air Force enlistees (Alley, Berberich, & Wilbourn, 1977; Alley, Wilbourn, & Berberich, 1976).

During VOICE development and subsequent validation, two types of scoring procedures were devised: (1) basic interest scales and (2) occupational scales. Basic interest scales were developed using factor analysis of individual item responses (Alley, Berberich, & Wilbourn, 1977). They are relatively homogeneous scales capturing 18 underlying dimensions of vocational interest without reference to particular occupational specialties. Basic interest scales range from 7 to 20 items per scale. Items are scored in a L-I-D (like, indifferent, dislike) format with values of 3, 2, and 1, respectively, and each basic interest score represents the summation of the item scores within a particular scale. Because standardized basic interest scale scores provide information concerning the extent to which an individual possesses interest in broad vocational areas relative to a normative population, these scores are useful even for respondents who might not have Air Force career aspirations.

The occupational scales were developed for specific Air Force occupational areas using empirical job satisfaction criteria. Twenty occupational scales were developed based on a DoD occupational classification. DoD occupational categories were used to allow conversion to occupational areas in other DoD agencies or the civilian sector. After subjects who had originally been administered the VOICE during Basic Military Training (BMT) were on the job approximately 1 year, they were sent a job satisfaction survey. Using a global job satisfaction criterion, common and gender-specific regression weights were developed for each of the basic interest scales within each of the occupational groups. Occupational scale scores were computed by multiplying the regression weights specific to the occupational area by their corresponding basic interest scale score, summing the cross products, and adding a constant (Alley, Wilbourn, & Berberich, 1976). The VOICE significantly predicted satisfaction in 15 DoD occupational groups representing 87% of all Air Force jobs.

Despite the success of the occupational scales, some problems needed to be overcome. Of particular concern was scoring complexity since operational constraints dictated the need for a quick hand-scoring capability. In addition, early briefings, concerning VOICE operational implementation, indicated that Air Force managers, accustomed to the four MAGE (mechanical, administrative, general, and electronic) aptitude scores, desired a reduction in the number of occupational scales. These concerns resulted in attempts to develop composites amenable to the operational environment.

METHOD

Development of VOICE operational composites was considered necessary in order to accomplish the following objectives: (1) to reduce the number of occupational scales and (2) to simplify scoring of the resulting scales.

Occupational Scale Reduction

In order to explore the possibility of reducing the 20 occupational scales to a more manageable subset, the original 20 occupational scales were factored using principal axis factor analysis with varimax rotation. A male standardization sample of approximately 10,035 cases was used for these analyses. Since this phase of operational composite development was exploratory, only a male sample was factored at this time. For the male sample, very little information loss occurred, as measured by the percent of trace, when the original 20 scales were reduced to eight composites. In the eight-factor solution, the percent of trace was 89.55. However, further reduction resulted in too much information loss, and in the four-factor solution, the percent of trace was 68.12. Although some managerial preference had been expressed during early VOICE implementation briefings for only four VOICE operational composites, factor analytic results dictated a minimum of eight composites. Apparently, the mechanical, administrative, and electronic areas can be reduced to a single interest score, as in aptitude assessment. However, it i. understandable that the G, or general, area does not lend itself to a single score predictive of a criterion since job content in this area is extremely heterogeneous.

In analyzing the factor loadings of the eight-factor solution, names were given to the factors on the basis of the original 20 scales. A separate mechanical factor emerged as did a separate administrative and electronics factor. The remaining five factors were considered subsets of the global general area and were labeled accordingly: G_1, Security and Support Services; G_2, Medical Care; G_3, Medical and Dental Technician; G_4, Utilities Maintenance, and G_5, Technical and Allied Specialties. These factors became, at least tentatively, the new VOICE operational composites. These composites and their factor loadings are presented in Table 1.

Scoring Simplifica*ion

Once the new composites were identified, it was necessary to develop an operational scoring procedure amenable to rapid scoring by hand. One obvious simplification could be introduced if composite scores were obtainable directly from item responses rather than through the process of first computing basic interest scores. The approach selected was based on the observation that the relative contribution of a single homogeneous interest scale to a composite is dependent on both the number of items in the scale and the magnitude of its regression coefficient. Also, the relative contribution of all subscales to a composite can be maintained to a reasonable degree under the constraint that each of the subscale regression weights are set equal to one, if the number of items within subscales are permitted to vary. This could be accomplished for any given composite by (a) retaining all items for that subscale with the highest absolute regression weight (i.e., set the regression weight to a standard value of 1.0) and (b) reducing the number of items considered in all remaining subscales by a proportionate amount. Since the scales are homogeneous, the items within a scale would be considered relatively interchangeable for purposes of selecting specific items to represent a scale.

As a basis for this procedure, each of the eight operational composites were used in turn as criteria in a regression analysis with the basic interest scales as predictors. The regression weights for each of the composites are presented in Table 2. The process of obtaining the requisite number of items from each subscale for the Medical Care (G_2) composite is illustrated in Table 3.

Using the method described above, sets of items were identified which could be used directly in computation of operational composites. Although items were considered relatively interchangeable due to their homogeneity, items chosen for the reduced sets were selected on the basis of high factor loadings on the original basic interest factors as well as similarity of loadings for males and females on these original basic interest factors. Thus, the homogeneity of items was maximized as was also the relevance of the items for both males and females.

Table 1

Factor Loadings for Rotated Factors, Eight-Factor Solution

	Operational Composite Factors							
	(M)*	(A)*	(G_1)	(G_2)*	(G_3)	(G_4)*	(G_5)*	(E)
Occupational Scale	Mech	Admin	Security & Support Service	Medical Care	Medical & Dental Tech	Utilities Maint	Tech & Allied Spec	Elect
1 Radio/Radar Equip Repair	.2087	-.1093	.1324	-.0262	-.0335	-.1431	.0290	.8165
2 Misc Elect Equip Repair	-.0223	.1233	-.0884	-.0486	-.1398	-.0137	.1067	.9308
3 Radar & Air Traf Control	.1571	-.3648	.5443	-.2002	.5197	.0765	.2296	-.2523
4 Misc Comm & Intell Spec	.5854	.0379	.1407	-.0161	-.3059	.2326	.2084	.5451
5 Medical Care	.0087	.0867	.2391	.9314	.1267	-.0934	-.0991	-.0536
6 Misc Med & Dental Spec	.1584	.1688	.0008	.1648	.8685	.1850	-.0058	-.1156
7 Tech & Allied Spec	-.0931	.0146	-.2005	-.0821	.0178	-.1416	.8943	.1227
8 Administration	.0712	.8777	.1188	.2180	.1794	.0949	.0030	-.1799
9 Misc Admin Spec & Clerks	-.0753	.8377	-.0418	-.1112	.0061	-.3193	-.0519	.1532
10 Gen Aircraft Mech	.8273	.1148	.2270	.1776	.2670	-.1997	-.2456	.1563
11 Aircraft Eng Mech	.8629	.1107	.0863	-.2109	-.0166	.0957	.1011	.0706
12 Aircraft Access Mech	.6479	-.3060	.1394	.0316	.2142	-.0192	-.1395	.5241
13 Armaments & Munitions	-.1050	.1406	.7502	.1030	.3195	-.0392	-.2435	.3149
14 Gen Mech	.6301	.1785	-.1339	.3885	.0783	.2858	-.0379	.4954
15 Utilities Maintenance	.0673	-.2049	-.3106	-.1091	.2593	.8197	-.2033	-.0927
16 Fire Fighters	.7019	-.3773	.2188	.2397	.1192	.0233	-.1730	-.3180
17 Material Rec, Stor, Issue	.1972	.5588	.5372	.4509	-.1287	.0955	.2203	.0350
18 Security Police	.3074	.0916	.7765	-.0240	-.1981	-.1412	-.3453	-.0050
19 Law Enforcement	.1326	.0042	.8768	.2961	.0472	-.1485	-.0649	-.1192
20 Misc Svcs & Supply	.4449	-.1326	.6124	.1065	-.3037	-.2581	.3296	.2736

*Denotes factor dimensions that have been reflected.

Table 2

Regression Weights for Basic Interest Scales on Operational Composites

	Operational Composite Factors							
	(M)	(A)	(G_1)	(G_2)	(G_3)	(G_4)	(G_5)	(E)
Basic Interest Scale	Mech	Admin	Security & Support Service	Medical Care	Medical & Dental Tech	Utilities Maint	Tech & Allied Spec	Elect
1 Office Admin	-.0349	.0678	.0732	.0159	.0457	-.0431	-.0013	.0041
2 Electronics	.0028	.0042	-.0063	.0216	-.0154	.0538	.0343	.0574
3 Heavy Construction	.0298	.0271	.0048	.0286	-.0099	.0313	.0481	-.0019
4 Science	-.0037	-.0339	.0251	.0188	.0461	-.0201	.0180	-.0130
5 Outdoors	.0616	-.0097	.0119	-.0451	.1239*	.0027	.0741	-.0562
6 Medical Service	-.0096	-.0072	-.0560	.0860	.0225	.0352	-.0584	-.0301
7 Aesthetics	-.0186	.0161	-.0310	.0195	-.0321	-.0114	.0131	-.0152
8 Mechanics	.0715	.0186	-.0440	-.0133	.0270	-.0247	-.0805	-.0300
9 Food Service	.0446	.0304	.0107	.0490	-.0173	-.0834	.0944*	.0210
10 Law Enforcement	.0394	-.0227	.0724	-.0136	-.0081	-.0852	-.0286	-.0109
11 Audiographics	-.0036	-.0387	-.0318	.0274	-.0897	-.0234	.0751	-.0601
12 Mathematics	-.0238	-.0356	-.0222	.0183	-.0154	.0061	-.0385	.0410
13 Agriculture	-.0337	-.0534	.0293	.0254	-.0581	.0223	-.0564	.0702
14 Teacher Counseling	.0158	.0483	.0246	-.0797	-.0278	.0388	.0231	-.0352
15 Marksman	-.1233*	-.0227	.0826	.0184	-.1156	.0013	-.0836	-.1190*
16 Craftsman	-.0765	-.0274	-.0505	-.1501*	-.0429	-.0744	-.0158	-.0555
17 Drafting	-.0207	-.0751*	.0166	-.1079	.0893	.0719	.0111	.0106
18 Auto Data Processing	.0014	.0441	-.1046*	-.0395	.0418	-.2202*	-.0426	.0871

*Largest weight given a proportion of 1.0 and used as devisor for subsequent calculation of proportions.

Table 3

Computational Example of Item Selection Procedure, Medical Care (G_2) Composite

Basic Interest Scale	Regression Weight	Proportion	Number Items (Full Set)	Number Items (Reduced Set)
1 Office Adm	.0159	.1059	20	2
2 Electronics	.0216	.1439	20	3
3 Heavy Const	.0288	.1919	20	4
4 Science	.0188	.1252	20	3
5 Outdoors	-.0451	.3005	15	5
6 Medical Ser	.0860	.5730	20	11
7 Aesthetics	.0195	.1299	15	2
8 Mechanics	-.0133	.0886	15	1
9 Food Service	.0490	.3264	15	5
10 Law Enforcement	-.0138	.0919	15	1
11 Audiographics	.0274	.1825	10	2
12 Mathematics	-.0183	.1219	12	1
13 Agriculture	.0254	.1692	15	3
14 Teacher/Counsel	-.0797	.5310	10	5
15 Marksman	.0184	.1226	7	1
16 Craftsman	-.1501	1.0000*	7	7
17 Draftsman	-.1079	.7189	7	5
18 Auto Data Process	-.0395	.2632	7	2

*Highest weighted value given a value of 1.0.

To determine the most efficient method of scoring the items within the constraints imposed by the operational environment, two methods of scoring actual item responses were explored: (1) the L-I-D method in which like, indifferent, and dislike options were scored and (2) the L-D method in which only like and dislike options were scored. Both methods involved simplified scoring methods without the need for computation of regression weights or sums of cross products which had been a requirement for occupational scale score computation.

L-I-D Method. The preferred approach from a psychometric perspective was to use all information available for each item by scoring responses as they were scored previously, using the L-I-D format scored 3, 2, 1, respectively. Regression coefficients were both positively and negatively weighted, thus specific item responses, regardless of the option selected, were considered either positive (added) or negative (subtracted) based on the sign of the regression coefficients associated with the basic interest scale from which they were derived. Templates were constructed for hand scoring purposes, and a program was developed for computer scoring in order that the scores could be validated against the original 20 occupational scales.

L-D Method. Although the L-I-D scoring method seemed best from a statistical viewpoint, it did not lend itself to extremely rapid hand scoring since the addition or subtraction of response options with different values was involved. The VOICE will need to be scored in Armed Forces Examining and Entrance Stations (AFEES), possibly in large volumes, either by hand or with equipment lacking sophisticated scoring capability. In order to facilitate scoring and devise a scoring method for use on equipment capable of distinguishing only dichotomous (binary) responses, a simpler L-D method was also developed which ignored indifferent responses and scored only like and dislike responses using a "hit" approach. Only addition was used, and positively valenced items were scored if a like (L) response was indicated; negatively valenced items were scored if a dislike (D) response was indicated for the item. For each composite the same items were used as with the L-I-D method.

RESULTS

In order to make a preliminary determination of the relative validity of these methods, an intercorrelation matrix was generated which correlated scale and composite score responses for all subjects in the AFHRL VOICE data files for whom predicted satisfaction data were available (N = 22,748; 12,713 females, 10,035 males). Means, standard deviations, and intercorrelations for the eight operational composites (scored using both L-I-D and L-D formats) and selected subsets of the 20 occupational scales for males, females, and total sample, respectively, are shown in Tables, 4, 5, and 6.

Generally, correlation coefficients between the eight operational composites and corresponding occupational scales were moderate to high, with higher coefficients being obtained for the L-I-D method than the L-D method, as would be expected. Except in the armaments and munitions occupational area, for which effective satisfaction prediction had been a problem in earlier research (Alley, Wilbourn, & Berberich, 1976), the trend toward

Table 4

Means, Standard Deviations, and Correlations for Occupational Scale Scores and Selected Operational Composite Scores, Male Sample

			Eight Operational Composites													
			(M)		(A)		(G_1)		(G_2)		(G_3)		(G_4)		(G_5)		E	
20 Occupational Scales			Mech		Admin		Sec & Support Service		Medical Care		Med & Dental Tech		Utilities Maint		Tech & Allied Spec		Elect	
			LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD
	$\bar{X}$		14	26	-20	32	18	30	11	19	30	28	-11	18	12	41	13	25
		SD	10	7	15	11	11	9	8	6	10	7	6	5	12	10	11	7
M_1 Gen Acft Mech	531	41	.76	.56														
2 Acft Eng Mech	582	60	.82	.62														
3 Acft Access Mech	518	56	.66	.48														
4 Armaments & Munitions	426	48	-.11	-.09														
5 Gen Mech	546	73	.51	.31														
A_1 Administration	536	37			.81	.49												
2 Misc Admin Spec & Clerks	579	44			.78	.53												
G_1 Radar & Air Traffic Con	594	72					.54	.28	.01	-.06	.35	.23	.03	-.06	.13	.04		
2 Misc Comm & Intell Spec	581	45					.26	.22	.03	.04	-.25	-.13	.05	.03	-.07	-.03		
3 Medical Care	606	64					.10	-.01	.75	.43	.03	-.04	.01	-.09	-.31	-.29		
4 Misc Med & Dental Spec	664	72					-.07	-.15	.09	-.01	.74	.42	.34	.14	.00	-.09		
5 Tech & Allied Spec	628	38					.08	-.04	-.10	-.16	.18	.07	-.05	-.13	.71	.34		
6 Utilities Mtce	632	67					-.30	-.09	-.11	.05	.17	.20	.70	.53	-.15	.02		
7 Fire Fighters	617	85					.23	.22	.27	.23	.13	.17	-.01	.00	-.16	-.06		
8 Material Rec, Stor, Issue	450	43					.34	.13	.37	.16	-.14	-.18	.10	-.04	.13	-.03		
9 Security Police	406	71					.72	.52	.01	.05	-.19	-.07	-.15	-.09	-.37	-.20		
10 Law Enforcement	520	102					.59	.40	.25	.19	.06	.06	.00	-.01	-.16	-.10		
11 Misc Svcs & Sply	506	63					.66	.37	.08	-.02	-.25	-.19	-.28	-.25	.30	.11		
E_1 Radio/Radar Equip Repair	627	38															.63	.40
2 Misc Elec Equip Repair	566	54															.75	.54

Table 5

Means, Standard Deviations, and Correlations for Occupational Scale Scores and Selected Operational Composite Scores, Female Sample

20 Occupational Scales			Eight Operational Composites															
			(M)		(A)		(G_1)		(G_2)		(G_3)		(G_4)		(G_5)		E	
			Mech		Admin		Sec & Support Service		Medical Care		Med & Dental Tech		Utilities Maint		Tech & Allied Spec		Elect	
			LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD
	$\bar{X}$		9	25	-18	35	16	30	13	21	31	30	-12	18	17	46	4	22
		SD	10	6	19	12	11	8	9	6	10	7	6	5	14	10	12	7
M_1 Gen Acft Mech	465	82	.54	.34														
2 Acft Eng Mech	485	144	.26	.20														
3 Acft Access Mech	458	108	.22	.14														
4 Armaments & Munitions (Males only)																		
5 Gen Mech	505	103	.07	.003														
A_1 Administration	561	75			.70	.57												
2 Misc Admin Spec & Clerks	566	49			.75	.60												
G_1 Radar & Air Traffic Con	610	150					-.05	-.08	-.25	-.22	.24	.15	.22	.10	.22	.11		
2 Misc Comm & Intell Spec	542	104					.13	.05	.03	-.04	-.21	-.20	.20	.06	.37	.16		
3 Medical Care	649	51					.32	.20	.40	.28	.41	.29	-.14	-.15	.30	.10		
4 Misc Med & Dental Spec	669	65					-.07	-.02	.18	.19	.71	.55	-.12	-.05	-.07	.00		
5 Tech & Allied Spec	642	68					-.32	-.21	-.18	-.15	-.16	-.12	-.20	-.11	.55	.39		
6 Utilities Mtce	537	196					.08	.04	-.01	-.04	.00	.01	-.12	-.13	.43	.27		
7 Fire Fighters (Males only)																		
8 Material Rec, Stor, Issue	467	77					.53	.38	.18	.13	.08	.06	.05	.01	.10	.07		
9 Security Police (Males only)																		
10 Law Enforcement	564	85					.58	.39	.17	.11	.08	.06	-.17	-.15	-.02	-.03		
11 Misc Svcs & Sply	463	97					.41	.25	.10	.02	-.24	-.19	-.21	-.23	.33	.15		
E_1 Radio/Radar Equip Repair	542	71															.59	.45
2 Misc Elec Equip Repair	519	76															.69	.49

Table 6

Means, Standard Deviations, and Correlations for Occupational Scale Scores and Selected Operational Composite Scores, Total Sample

20 Occupational Scales			Eight Operational Composites															
			(M)		(A)		(G_1)		(G_2)		(G_3)		(G_4)		(G_5)		E	
			Mech		Admin		Sec & Support Service		Medical Care		Med & Dental Tech		Utili-ties Maint		Tech & Allied Spec		Elect	
			LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD	LID	LD
	$\bar{X}$		11	25	-18	34	17	30	12	21	30	29	-10	18	15	44	8	23
		SD	10	7	18	12	11	8	9	6	10	7	6	5	13	10	13	7
M_1 Gen Acft Mech	492	63	.71	.45														
2 Acft Eng Mech	525	93	.72	.49														
3 Acft Access Mech	487	73	.59	.38														
4 Armaments & Munitions	426	48	-.11	-.09	(Males only)													
5 Gen Mech	518	73	.56	.32														
A_1 Administration	548	57			.76	.59												
2 Misc Admin Spec & Clerks	571	44			.83	.62												
G_1 Radar & Air Traffic Con	581	65					.49	.29	-.19	-.19	.34	.22	.08	-.01	.18	.08		
2 Misc Comm & Intell Spec	553	63					.23	.10	-.07	-.13	-.29	-.29	.22	.06	.02	-.11		
3 Medical Care	619	52					.11	.02	.84	.56	.28	.18	-.07	-.13	-.05	-.08		
4 Misc Med & Dental Spec	667	61					-.06	-.06	.17	.14	.81	.57	.13	.08	-.04	-.03		
5 Tech & Allied Spec	640	39					-.05	-.04	-.10	-.09	.02	.03	-.23	-.17	.87	.57		
6 Utilities Mtce	577	76					.07	.02	-.33	-.28	-.10	-.11	.49	.29	-.04	-.09		
7 Fire Fighters (Males only)	618	85					.23	.22	.27	.23	.13	.17	-.01	.00	-.16	-.06		
8 Material Rec, Stor, Issue	454	46					.19	.28	.31	.17	-.01	-.05	.06	-.04	.21	.07		
9 Security Police (Males only)	406	71					.72	.52	.01	.05	-.19	-.07	-.15	-.09	-.37	-.20		
10 Law Enforcement	540	77					.79	.50	.25	.15	.11	.07	-.12	-.13	-.06	-.07		
11 Misc Svcs & Sply	454	46					.60	.35	.06	-.05	-.31	-.27	-.19	-.24	.25	.05		
E_1 Radio/Radar Equip Repair	586	54															.86	.60
2 Misc Elec Equip Repair	550	55															.89	.65

moderate to high coefficients for corresponding scale and composite scores was evident for mechanical, administrative, and electronic areas. However, due to the heterogeneity of the general category, such a trend was less evident when general scale scores were correlated with general composite scores, except where specific comparisons were made. For example, scores on the security police scale were relatively uncorrelated with the medical care composite scores. However, the correlation of scale scores with like-named composite scores tended to be positive with moderate to high values, except in the case of utilities maintenance for female subjects.

It is interesting that even the validity coefficients for the female sample were relatively high despite the fact that the eight composites were based on a factor solution applied to a male sample. However, this fact might in part be explained by considering the similiarity between factor loadings for males and females on items selected for the operational scoring composites. Also, the original factor structure for the basic interest factors was similar for males and females and, in all but a few cases, common rather than gender-specific regression weights were used in computing the occupational scale scores. Thus, the ability of the occupational composite scoring procedures described in this paper to hold up well when applied to a female population is understandable, despite the fact that a male population was used for initial model building.

DISCUSSION

Results of this study indicated that the methods developed for operationalizing the VOICE show promise. One of the most interesting of current findings was the degree of redundancy which apparently existed in the original 20 occupational scales. In the mechanical, administrative, and electronic areas, the original subcategories were not substantially different with respect to interests since the original scales in each of these areas were able to be reduced to single composites without considerable loss of information. Also, a factor structure emerged which was consistent with the existing DoD aptitude categories with the exception of the heterogeneous general area. However, the factor structure identified was based on only a male sample, and prior to operational implementation of the composites, it will need to be determined if the same or a similar factor structure will emerge if an eight-factor solution is applied to a female sample and if little information loss will occur in such a solution.

Of the two operational scoring methods, the L-I-D method was stronger from a psychometric viewpoint than the L-D method, as evidenced by the typically higher correlations with selected occupational scales and the fact that no information was ignored. From this perspective, it is the preferred method for operational use, and at this stage of development the L-D method can be recommended for use only if large scale processing is required and available scoring equipment necessitates tabulation of binary responses. Although the L-I-D method is preferred, it does require more time to score manually, and future efforts might be directed at simplifying the scoring of this method, while retaining the three-option format. The L-D method might also be refined through further research. Loss of information resulted when L-D scores were computed since indifferent scores were ignored. This fact might be partly responsible for the decrement in correlation coefficients relative to the L-I-D method. It will need to be determined if scores computed on a new sample for which the indifferent response option was eliminated, rather than just ignored, will result in more robust coefficients. Also, using existing data, methods of scoring the L-D format in which indifferent responses are scored as if they were like or dislike responses need to be explored. The validities of these alternative methods will need to be determined and compared with the validity coefficients already obtained. As an adjunct to these efforts, a determination will need to be made concerning the relative merit of another type of operational scoring method currently under development but intended specifically for machine processing: adaptive VOICE interest testing (Watson, 1979).

The technique used to derive composites for VOICE operational use may have general applicability to other situations where homogeneous scales are involved, and use of only a portion of original scale items along with scoring simplicity is desired. One of the most appealing features of the approach described in this paper is the elimination of intermediate steps, allowing computation of composite scores from items directly.

Once these operational methods have been refined and a final method, or methods, selected for implementation, VOICE data can serve as an effective and efficient tool for measuring the vocational interests of Air Force personnel. Used in conjunction with aptitude data and other information, it should enhance the ability of the Air Force to make sound job placement decisions by systematically relating expressed interests, prior to entry into a job, to subsequent job satisfaction in a variety of occupational areas. Neither knowledge of Air Force jobs nor vocational sophistication is necessary.

Eventual use of the VOICE is not limited solely to use with Air Force recruits. The VOICE is currently being used on a test basis for reenlistment counseling and cross-training purposes with Air Force job incumbents. It could also be used with other military and civilian populations provided it is normed and validated on the populations to which it might be applied.

REFERENCES

Alley, W.E., Berberich, G.L., & Wilbourn, J.M. Development of factor-referenced subscales for the Vocational Interest-Career Examination. AFHRL-TR-76-88, AD-A046 064. Brooks AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, June 1977.

Alley, W.E., Wilbourn, J.M., & Berberich, G.L. Relationships between performance on the Vocational Interest-Career Examination and reported job satisfaction. AFHRL-TR-76-89, AD-A040 754. Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1976.

Echternacht, G.T., Reilly, R.R., & McCaffrey, P.J. Development and validity of a vocational and occupational interest inventory. AFHRL-TR-73-38, AD-774 573. Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1973.

Ward, J.H., Jr., Haney, D.L., Hendrix, W.H., & Pina, M., Jr. Assignment procedures in the Air Force Procurement Management Information System. AFHRL-TR-78-30, AD-A056 531. Brooks AFB TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, July 1978.

Watson, T.W. Computerized vocational counseling using the Vocational Interest-Career Examination (VOICE). Paper presented at the 21st annual conference of the Military Testing Association, San Diego, October 1979.

STANDARDS: FROM IVORY TOWER TO REAL WORLD

William A. Gager, Jr., Ph.D.

Chief of Naval Education and Training
Pensacola, Florida 32508

Perspective

I would ask you to consider the role of standards for assessing the achievement of learning objectives, and their use for external evaluation in a large operational training system. As points of reference of what we have enviously described as "Ivory Tower," let me contrast three se^tings: (1) standards used by an instructional system developer to plan the assessment of a student's achievement at the end of a course module in a school, (2) standards used by a researcher in a funded project of many months duration treating a limited number of courses or skill areas to develop a high quality performance testing system, and (3) the routinized external sampling of product of a large, cost-conscious, undermanned training system to determine if successful graduates have the job entry skills and knowledge which are the objectives of the various training courses. ("F^ternal" is used to mean after the graduate has departed the locatio.ı and jurisdiction of the training agency.) My remarks will focus on the thi^d setting: external appraisal in a large scale operating setting. If I may assume you know a great deal about the first two settings, I will describe my part of the "real world."

Thesis: Indirect and Echeloned Use of Standards

My thesis will be that for ongoing external appraisal in an austerely funded large scale training agency, standards of performance for assessing graduate achievement of job entry skills and knowledge will often be used indirectly in the feedback collection process, or will most commonly be applied in a restricted or abbreviated way in an echeloned system. I will cite minor exceptions in just a moment, will further explain the idea of indirect and echeloned use of standards, and comment on the implications of these notions for implementation.

The Organizational Context

"Generalizing from a sample of one," it will be assumed that certain characteristics of the Naval Education and Training Command (NAVEDTRACOM) are common in large organizations. The three salient ones are:

1. Size. Feedback is needed on about 3100 unique courses of instruction (which might more generally be called "programs"). If each program were reviewed every two years, 30 assessments would have to be completed each week. The requirement for a fast, inexpensive, adequately reliable method is obvious.

2. Noninterference with the Users of the Graduates. The employer (in our case, the Fleet) desires high quality graduates but is unable to take large

amounts of time away from the primary mission to assess the adequacy of training. Fleet units are also unable to dedicate extensive administrative support. This dictates several things:

a. Minimum adequate sample sizes

b. Simplicity of administration of routine monitoring

c. Short surveys

d. The policy that no information will be sought from the fleet which can be generated within the Training Command. This point will be important in considering indirect application of standards.

e. Terminology and format which can be used by nonprofessionals

3. Staff and Dollar Costs. The mere size of the system makes cost critical. The NAVEDTRACOM is austerely manned. (This is bureaucratic language for understaffed.) There are bitter competitions for existing funds between developments to meet new requirements, applications of modern instructional technologies, and new support equipment to match fleet procurements. In the fleet there are competing demands for operations, maintenance, continuing on-board training, and crew morale. The resources for assessment must be in balance.

Purposes of External Evaluation

Since the perspective of our focus is external evaluation, we must ask what its purpose is in order to consider whether standards are essential. In the NAVEDTRACOM, the basic purpose of external evaluation is to establish whether a problem exists in terms of graduates having designated job entry skills and knowledge at the time of reporting; and whether the objectives of the course instruction remain the priority needs of the fleet. It is, of course, implied that the Training Command should then be able to react to the feedback findings developed.

Implications for Implementation of a System

When the reasons of costs, time requirements, low-impact on the fleet resulting from large scale administrations, and practical feasibility in view of competing demands were compared to the essential purpose of a feedback system, the NAVEDTRACOM management decision was to use a survey system for the basic routine monitoring function. The surveys were to repeat the items on a "Skills Profile" of each course. The Skills Profile was to contain the job entry skills, task competencies, or knowledge which the graduate should have upon reporting to his utilization assignment. Items were to correspond to terminal objectives of the course of instruction and were to be at a level of generality to comprehensively cover the course in a maximum of 60 items. The survey system is labeled "Level II Feedback," Level I being unsolicited feedback, or feedback from existing reports primarily used for another purpose.

Several factors emerged during initial development:

1. Standards on rate of production would be required on skills such as typing or sending and receiving code.

2. For a very large proportion of objectives, standards were implied and were either "100% of the time," or a correct act or piece of knowledge was required which was "go-no go."

3. While there are certain commonalities, in some cases standards of performance in the same rate and rating on a given skill differ because of differences between ship types, equipment, manning patterns, and specific work assignments.

4. In many cases it appeared likely that frequently at the level of generality of the Skills Profile, less than adequate performance would be more related to substantive inadequacy on a discrete element of a task or a subordinate piece of knowledge than to level of performance per se.

5. No standard is appropriate in the selected type of feedback collection when the measurement involved is not a part of normal operations.

6. Since supervisors are a critical given part of the manpower utilization system, in the absence of a major intervention to change the relationship decision makers must have appropriate trust and confidence in the training and common experience of these supervisors, and acceptance of the standards implied by their common experience.

7. In many cases, standards have not been developed or negotiated between the NAVEDTRACOM and the sponsor of a given course.

8. Standards of performance are in many cases contained in other technical documentation which describes equipment, procedures, or duties of watch stations.

9. If there is a difference in an individual case between standards developed for course development and standards required by a specific duty setting which a graduate goes to, feedback on adequacy of training should be based on the actual job requirements of the job being reported.

In view of these factors, the standards used in the initial implementation of the NAVEDTRACOM external appraisal system are highly abbreviated or implied standards based on both technical publications and standard practice are assumed.

No suggestion is being made that standards of performance are not critical to design of training and internal evaluation of instructional effectiveness. Rather the issue is what standards are essential to meet the purpose of the feedback collection.

The appropriate use of standards will be the topic of continuing assessment and modification of the system. The "real world" judgment is that modification of abbreviated standards, implied standards, "go-no go" standards, and standards based on common publications and practice averaged across a sample will not significantly change the basic findings of a feedback system until sample-based and negotiated standards are developed as each course is revised using the best practice of instructional system development.

Turning to indirect use of standards in collecting feedback by survey, the questionnaire usually does not include the specific option of asking whether the respondant agrees with the standards set by the course developers and trainers. It is therefore useful to establish a link between feedback findings and standards used in training, whether stated or unstated. Each Skills Profile item must be linked directly to a module of the course. The feedback finding in the NAVEDTRACOM system not only identifies a terminal skill or knowledge which is perceived to be inadequately trained to, it produces a list of the graduates who were reported to be inadequately trained. Graduates reported as unsatisfactory can then be compared to their rating on the related end-of-module test or performance check, or to the length of time required for completion. In the case where a predominance of the graduates reported as unsatisfactory by the fleet showed no indication of inadequacy at the time of completion of the module, there is the strong suggestion that the standard was set too low or too differently from the fleet needs.

From the standpoint of external appraisal, we issue the call for trainers to keep records and develop procedures so that on the problem items identified by feedback a comparison can be made by individual between the degree of competency demonstrated on end-of-module checks and the "adequacy" assessed in the fleet.

An additional extension of this approach is to make comparisons between unsatisfactory feedback findings and the ratings which the specified graduates made on individual test items of end of module tests or performance checks. When the unsatisfactory graduates on a given Skills Profile item are found to have in common that they were rated as unsatisfactory on a particular end of module test item, the more general feedback findings can be related to a specific weakness.

Objectivity/Subjectivity

A comment may be in order on objectivity and assessment of standards of performance. Observers will note that the NAVEDTRACOM Feedback and Appraisal System Level II Surveys collect subjective data based on observed performance and related indicators. No one I know will argue that subjective feedback is more reliable than objective feedback. But many managers will argue two points: (1) The difference between objectivity and subjectivity is not a dichotomy. It is a continuum within which variability associated with subjectivity can be reduced. The subjective report of one leading chief about the ability of a recent graduate to perform is certainly less reliable than the pooled subjective assessments of a very large sample of leading chiefs. The same could be said about establishing standards about what is expected by a supervisor. Pooled over a large sample, a mean expectation or common expectation tends to emerge. (2) The second point managers will argue is that if there are limiting constraints relating to dollars, time, or tolerable disruption, high quality broadly based subjective information is better than no information. A potential implication of this position is that if there are significant economies of time and dollars in high quality subjective methods over objective method which would make the subjective method feasible for implementation in the large scale monitoring role, then developmental effort should be directed toward improvement of quality of subjective methods.

An Echeloned External Appraisal System

When we cast three settings in the opening remarks, I did not elaborate on one important step in the external appraisal setting, namely: What do you do when a problem appears to exist, you have gone as far as you can with the information collected by monitoring "customer satisfaction" based on terminal objectives of a course, and you still can't identify the specific problem or its cause? It appears clear you turn to more sophisticated, more discriminating, perhaps more reliable, and usually more expensive methods. It is at this stage that detailed standards come into use in performance testings, structured interviews, or more detailed surveys. In the Navy Training Command System this is called "Level III Analysis." (You will recall surveys of supervisors of recent graduates are called "Level II Feedback," and unsolicited feedback is called "Level I Feedback.") What does an echeloned system accomplish? It screens out the lesser problems which can be recognized and addressed directly; and focuses the scarce resources for detailed analysis where confirmed problems exist. On the special problems the performance deviations can be measured and their causes analyzed. The more rigorous procedures developed by the researchers can be selectively put into service in relation to the resources available to the Training Command.

Summary

I have argued that standards cannot in all likelihood be applied to routine large-scale external appraisal directly as they were developed to support course development and internal measurement of achievement. It has been suggested that we should be prepared to apply the details of standards in echelon ranging from implied standards or abbreviated standards for many items in routine surveys to more and more specific standards for detailed analysis of specified selected problems. It has been further suggested that we should be getting procedures and records in place to link feedback findings based on course objectives to more detailed internal end-of-module test and performance check results, by individual, in order to attempt to see what the graduates who were reported to appear unsatisfactorily trained had in common while they were in school.

We agree that standards of performance must exist, but like the good silverware, it may not be appropriate to use it for every meal.

STANDARDS DERIVED FROM A BASIC EQUIPMENT SKILLS AND TRAINING MATRIX

Gerald J. Laabs

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

Problem

The Navy training community and the Fleet often have difficulty in communicating about training problems because Fleet performance cannot be adequately related to course content. One problem is that the Fleet has limited information on course content. Consequently, the Fleet does not know what a school graduate is supposed to be able to do after arriving onboard ship, and the Fleet also may have difficulty in deciding which training course an individual should attend. The Navy Training Command's concept of a Skills Profile may help solve this problem.

Another problem is that the training community receives limited, subjective feedback on school graduates because the Navy does not have a comprehensive, standardized procedure for objectively measuring job performance. Our Center is developing a prototype Performance Proficiency Assessment System which may help solve this problem. This system, which corresponds to the Level III analysis of the Navy Training Command's Training Appraisal Program, is being patterned after industrial quality control systems in which standards are established, samples of a product are measured, the discrepancies between the standard and the measurement are traced back to their source, and corrective actions are taken. The purposes of the system are: (1) to measure the performance capabilities of samples of members of a rating in critical tasks areas (i.e., performance domains), and (2) to present analyzed performance data to personnel managers in a form that will help them initiate corrective actions.

Purpose

Without a framework to link job performance to what is being taught in the various courses, the replacement of subjective evaluation with objective performance tests in our measurement system will still not provide the training community with adequate feedback. The purpose of this effort was to develop a methodology to define performance domains for testing in the Performance Proficiency Assessment System and to provide a framework to relate job performance data to course content.

GENERAL APPROACH

Two obvious ingredients of a framework to relate job performance and training courses are a taxonomy of skills to form a basis of the measurement program, and detailed training course information. A third ingredient is equipment because the Fleet is concerned with skills related to specific equipment, and the content of courses is determined primarily by the equipments.

In the first phase of the effort, procedures were developed to define three elements--equipments, skills and training course information--for the Interior Communications Electrician (IC) rating. These elements were then combined using a matrix technique to produce a Basic Equipment Skills and Training Matrix. In the second phase, additional procedures were developed to: (1) produce the matrix by-product of a list of performance domains for the IC rating, and (2) to select the critical performance domains for testing. Test construction, data collection, and the provision of feedback remain to be done.

PHASE I: MATRIX DEVELOPMENT

The definition of the matrix elements during Phase I of the effort involved: (1) identification of major IC system and equipment clusters, (2) selection of a comprehensive set of job-task performance categories for the IC rating, and (3) preliminary classification and consolidation of the learning objectives from all of the major courses that members of the IC rating routinely attend. The equipment and skill elements were arranged as opposing dimensions in a two-dimensional matrix, then the consolidated learning objectives were entered into the cells of the matrix to produce the Basic Equipment Skills and Training Matrix.

System/Equipment Dimension

One dimension of the basic matrix was formed by 22 IC system and equipment clusters taken from the Equipment Identification Code Manual. These clusters, which fall under six superordinate headings, were further divided into subcategories to facilitate the eventual detailed classification of the learning objectives. Table 1 shows the IC system and equipment categories and superordinate headings as a dimension of the matrix within which learning objectives were categorized.

Skills Dimension

The second dimension of the basic matrix was formed by seven job-task performance categories taken from the Navy Occupational Task Analysis Program (NOTAP). The job-task analyses conducted by NOTAP specify what tasks a rating does, and result in a task inventory comprised of specific performance statements related to broad categories of equipment. In a recent study, Powers (1977) demonstrated the communality of some of the action verbs of these statements for 30 technical ratings. Virtually all personnel in almost every technical rating reported that they performed tasks in each of the following seven generic job-task categories: (1) Assemble/Disassemble, (2) Test/Inspect, (3) Troubleshoot/Repair, (4) Clean/Lubricate, (5) Adjust/Align, (6) Remove/Replace, and (7) Operate/Secures. In the case of the IC rating, all 20 respondents in Power's study indicated involvement in each of the seven categories. Table 1 shows the job-task categories as a dimension of the matrix within which learning objectives were categorized.

Table 1

Basic Equipment Skills and Training Matrix

JOB TASK CATEGORIES	ELECTRIC POWER GENERATION/DISTRIBUTION AND RELATED SYSTEMS				PROPULSION SYSTEMS		NAVIGATION SYSTEMS-ELECTRONIC/NON-ELECTRONIC					INTERIOR COMMUNICATION SYSTEMS						AUXILIARY SYSTEMS			STANDARD & SPECIAL EQUIPMENT	
SYSTEMS/EQUIPMENTS	ELECTRIC POWER GENERATING SYSTEMS	ELECTRIC POWER DISTRIBUTION SYSTEMS	STATIC POWER SUPPLIES	MISCELLANEOUS SM DISTRIBUTION SYSTEMS	MAIN STEAM PLANT APS-EQUIPPED	MAIN STEAM PLANT EOS-EQUIPPED	GYROCOMPASS SYSTEMS	UNDERWATER ELECTRIC MAGNETIC LOG SYSTEMS	DEAD RECKONING SYSTEMS	MAGNETIC COMPASS SYSTEMS	OPTICAL AND MISC NAVIGATIONAL AIDS	TELEVISION SYSTEMS GENERAL	AMPLIFIED VOICE COMMUNICATION SYSTEMS	TELEPHONE SYSTEMS	SAFETY AND WARNING ALARM SYSTEMS	SHIP ORDER AND INDICATING SYSTEM	RECORDING AND PROJECTION SYSTEMS	STEERING AND SHIP'S CONTROL SYSTEMS	FIN STABILIZING SYSTEMS	BALLAST & STABILIZING SYSTEMS	ELECTRICAL ELECTRONIC TEST EQUIPMENT	HAND TOLLS MANUALLY OPERATED EQUIPMENTS
ASSEMBLE DISASSEMBLE	3	3		2	1	3	17	3	3	3	3			7		3		3				
TEST INSPECT	4	21	3	1	2	4	79	9	28	3	7		20	13	5	10	2	3	6			
TROUBLESHOOT REPAIR	6	10		4	7	17	58	7	36	3	4	1	48	22	22	6	5	8	2	1		
CLEAN LUBRICATE	2	2	1				11	4	3		2		6	1	1				2			
ADJUST ALIGN	2	5	1	2	1	10	86	13	55	11	11		2	10	5	19	1	2	8		7	
REMOVE REPLACE		2	1	2			41		4				4	8	1				5		15	
OPERATE SECURE		1			23	26	11	4	13	2	5		14	12	4	10	1	2	2		15	2

Learning Objectives

Over 2500 objectives were extracted from 28 IC training courses. The first step in the process was to classify the objectives into one of three preliminary categories: job-task performance objectives, supporting skill/knowledge objectives, or administrative and military objectives. Only the job-task performance category is of interest for this paper. This category consisted of objectives that required a "hands-on" approach to accomplish the assignment such as those governing laboratory exercises or the completion of job sheets in which the trainee actively performs a sequence of steps.

Following the extraction of the objectives, they were consolidated by combining separate but related objectives into a single statement. All of the objectives were serialized using a six-position alphanumeric symbol. The digits in the first two positions refer to the training course. The course designation field is followed by the letter "P" to indicate that the objective is a job-task performance objective. The remaining three positions in the field are used to serially number the objectives as they are entered into the matrix.

Table 1 shows the number of job-task objectives entered into the matrix defined by the dimensions of IC Systems/Equipments and Job-Task Categories. A dictionary was also developed that listed the alphanumeric symbols of the objectives and either the course lesson or job sheet number from it was extracted.

PHASE II: PERFORMANCE DOMAIN DEFINITION

Although our measurement system limits testing to relatively small samples from the IC rating, it is obvious that it is not feasible to performance test each of the objectives entered into the Basic Equipment Skills and Training Matrix. During Phase II of the effort, the objectives were collapsed into task areas or performance domains, and then the most critical of these domains were selected to make our testing program manageable. Before the derivation of the performance domains from the Basic Equipment Skills and Training Matrix is described, a couple of other potential matrix by-products should be mentioned. For example, a Skills Profile for a course could easily be developed by listing the generic job-task action verbs and the corresponding systems, equipments, components, or component parts that the learning objectives in the matrix indicate are the objects of these verbs. This listing would tell the Fleet what a school graduate is supposed to be able to do after arriving onboard ship. Another useful by-product could be easily derived by entering course identification numbers into the matrix, rather than all of the objectives. Such a matrix could serve as a course selection guide for the Fleet and, along with course Skills Profiles, aid in the decision of which training course an individual should attend to learn specific equipment skills.

Deriving Performance Domains

The derivation of a list of performance domains for the IC rating is also relatively simple using the matrix. The sole reliance on learning objectives to define the domains, however, excludes those domains that are covered by on-the-job training rather than classroom instruction. To ensure that all domains were considered, 314 task inventory statements from NOTAP that referenced job performance were entered into the matrix along with the learning objectives. An interesting result of augmenting the matrix in this way was the identification of job tasks taught in school but not included in the NOTAP inventory, as well as those included in the NOTAP inventory but not taught in school.

Performance domains were formulated from the augmented matrix by dropping those cells that had no entries and clustering the remaining cells under higher-order task areas whenever possible. The action verb used for the performance domains defined by cells that could not easily be clustered with any others reflected the generic job-task category of the cells, either by repeating the category or substituting a single word such as "calibrate" for assemble/disassemble. The action verbs used for the performance domains defined by clustered cells were systematically determined by the generic job-task categories involved: "maintain" was substituted for the cluster of all categories including operate/secure if this was required for a maintenance action; "perform planned maintenance" was substituted for the cluster of test/inspect, clean/lubricate, and adjust/align; "perform preventive maintenance" was substituted for the cluster of test/inspect and clean/lubricate; "troubleshoot" was substituted for the cluster of troubleshoot/repair and remove/replace, and "teardown" was substituted for the cluster of assemble/disassemble and remove/replace. A combination of single action verbs were used for the performance domains defined by all other clusters of cells, such as "test and operate" for test/inspect and operate/secure. A subject matter expert (i.e., an IC Chief Petty Officer) further refined the clustering of the tasks. The final list consisted of 63 performance domains and a brief description of the content of each domain, as shown in the following examples:

TROUBLESHOOT ALARMS

Troubleshoot alarm panels and alarm/indicating system, replace components

MAINTAIN ELECTRONIC INTERCOMMUNICATING SYSTEMS

Establish and control communications, perform preventive and corrective maintenance

Critical Task Identification

The 63 performance domains and their brief descriptions were placed individually on computer cards so that a card-sort, based on the Q-sort methodology, could be carried out. For this card sort, 24 IC instructors from the IC School in Great Lakes were asked to look at the performance domains and decide on their relative importance with respect to having a good IC gang. First, the judges were given a shuffled deck of cards and told to place approximately one-third of them into each of three stacks: one for relatively important domains, one for relatively unimportant domains, and one for domains that could not be classified easily. Next, they were told to sort the three stacks into nine categories ranging from (1) Relatively Unimportant to (9) Relatively Important. The number of cards to be placed in each category was chosen so that the final distribution would be approximately normal.

Table 3 shows the top 15 performance domains yielded by the card-sort. An intraclass correlation coefficient derived from a single factor, repeated measures analysis of variance provided a measure of interrater agreement on the rank ordering of the domains. This coefficient, which was .35, is an estimate of the average reliability of a single judge for the card-sort. This reliability was entered into the Spearman-Brown prediction formula to estimate the reliability of the mean of the rank orders obtained from the 24 judges, which was .93. Thus, we can be reasonably sure that testing the top 10 to 15 performance domains will yield data on critical job tasks.

Table 3

Top 15 Performance Domains
From a Card-Sort

Rank	Domains
1	Operate and Troubleshoot Electrical Gyrocompass
2	Use Standard and Special Test Equipment
3	Troubleshoot Alarms
4	Troubleshoot Announcing Systems
5	Troubleshoot Ship's Metering and Indicating Systems
6	Troubleshoot Control Consoles
7	Use Written Materials in Support of Maintenance
8	Operate and Calibrate Alarm Systems
9	Maintain Components of Ship's Control Systems
10	Maintain Announcing Systems
11	Perform Preventive Maintenance on Ship's Metering and Indicating Systems
12	Perform Planned Maintenance on Electric Gyrocompass
13	Troubleshoot Sound Powered Telephones
14	Maintain Electric Intercommunicating Systems
15	Maintain Electronic Intercommunicating Systems

FOLLOW-ON

The development of the Basic Equipment Skills and Training Matrix for the IC rating allowed the easy derivation of performance domains that can be defined in terms of learning objectives. This feature of the performance domains, along with the matrix and dictionary of objectives, provides the framework to relate job performance data to course content. To indicate how this framework can be applied, the three major research phases that still need to be completed to fully construct and exercise the quality control measurement system will be outlined. These phases are: (1) seek existing sources of performance data and build performance tests where needed, (2) collect performance data from samples of members of the IC rating, and (3) feed back the analyzed performance data to the training community.

The gathering of performance data from existing sources when available, along with selecting only critical performance domains for testing and only testing samples of a rating, are factors that we hope will enhance our chances of developing a feasible and cost-effective Performance Proficiency Assessment System. When performance data are not available from existing sources, however, performance tests will have to be constructed. To identify the tasks that are candidates for a performance test, we simply reverse the procedures used to derive the performance domain. Using the performance domain of "Operate and Calibrate Alarm Systems" as an example, the first step is to examine the learning objectives that describe the tasks making up the domain. In this case there are 29 objectives from 3 different courses. Table 4 shows an example of some of the objectives included in this domain.

Table 4

Example Learning Objectives for the Domain
Of "Operate and Calibrate Alarm Systems"

Identification Symbol	Objective
14P017	Adjust/Align meter check point on the salinity indicating system (1SB).
14P018	Adjust/align alarm set points of each cell in the 1SB.
14P019	Adjust/align flasher contacts on salinity indicating system (1SB).
14P021	Operate/secure the wrong direction alarm system (DW).
14P022	Adjust/align throttle valve contacts of wrong direction alarm system.
01P036	Operate/secure the salinity indicating system (SB).
02P038	Adjust/align the alarm setpoint of each detector on the temperature monitor system.

The total number of objectives chosen for testing will depend upon the time and resources available for the test, but at least one objective will be chosen for each of the generic job-task categories represented in the domain. If objective 14P022 in T-ble 4 is one of those chosen, the next step is to consult the learning objective dictionary to find the course lesson or job sheet from which it was extracted. In this case, the objective was extracted from Job Sheet 9.4.1 in the Propulsion Alarms and Indicating Systems Maintenance Course. Table 5 shows this job sheet. Note that it contains the step-by-step procedures to complete the task specified in the objective and can easily be converted to an objective check-off sheet to be used in observing the task. Some job sheets and laboratory exercises are not as structured as this and require subject-matter expertise to develop the step-by-step procedures for the performance test. Subject-matter expertise is also needed to combine procedures from several objectives included in a domain into one test problem.

Table 5

Excerpt From a Job Sheet From the Propulsion Alarms
And Indicating Systems Mainenance Course

JOB SHEET 9.4.1	Circuit (DW) Wrong Direction Alarm System "Adjustment of Throttle Valve Contacts"

The following is a logical sequence to follow to adjust the contact makers on the ahead and astern throttle valves.

1. Secure power to circuit (DW) wrong direction alarm system, and tag out of service.

2. Insure both ahead and astern throttle valves are closed (fully clockwise).

3. Remove cover on the ahead throttle valve contact maker.

   A. Located in recess on propulsion panel, behind ahead throttle valve.

   B. Remove one lead from terminal strip.

4. Using PSM-4 Multimeter on the RX1 Scale.

   A. Connect meter across switch inside contact maker. (One lead to lead removed from the terminal strip, and the other lead connected to the terminal strip.)

.

.

.

13. Reconnect swith lead to terminal strip and replace cover on the contact maker.

14. To adjust the astern contact maker follow the above procedure, substituting the word astern for ahead in the steps of procedure.

15. After completion of adjustment of both the ahead and astern contacts an operational test is required to insure proper system operation.

SEE INSTRUCTOR FOR FURTHER INSTRUCTIONS.

The data collection phase involves gathering data from performance test administration and from existing sources. An example of data from an existing source is how well personnel performed the actions specified on maintenance requirement cards which are routinely observed in the Planned Maintenance System Inspection. Once all of the data have been collected, they must be analyzed and fed back to the training community. At present, we anticipate that such feedback might take the form of a report of the percent of a sample who could not accomplish the task, accompanied by a tally of the particular steps in the procedures observed that caused the failures. Because of the existence of the Basic Equipment Skills and Training Matrix and the learning objective dictionary, the analyzed performance data can be referenced to a specific course, and to one or more learning objectives that are covered in a specific course lesson or job sheet. The provision of feedback to the training community completes the feedback loop.

In summary, the Performance Proficiency Assessment System described above is a quality control measurement system in which (1) standards and performance tests for a rating are derived from learning objectives through the use of a Basic Equipment Skills and Training Matrix, (2) samples of members of a rating are measured, and (3) the discrepancies between the performance and the standards are reported to the training community to help them initiate corrective actions.

REFERENCE

Powers, T. E. Selecting presentation modes according to personnel characteristics and the nature of job tasks. Part I: Job tasks (Navy Technical Information Presentation Program Report). Bethesda: David W. Taylor Naval Ship Research and Development Center, January 1977.

The opinions expressed in this paper are those of the author and not necessarily those of the Navy Department.

CRITERION vs NORM REFERENCED MEASURES

Dorothy von K. Scanland, Ed.D.

Defense Activity for Non-Traditional
Education Support
Pensacola, Florida 32509

So much has been spoken and written concerning the pros and cons of criterion and norm referenced measures that I hesitated to offer another discussion on this topic. Yet it seems to me that any forum on standards for measuring performance and objectives achievement has to contain considerations of criterion and norm referenced measures. On that premise, then, I shall add my two cents worth.

I shall not attempt to pit criterion against norm referenced measures in an effort to prove that one is superior to the other, but try to review the rationale, indicate the likely need for each under differing circumstances and requirements for measuring, and point out the kinds of uses best served by each.

Because norm referenced measures have been so popular these past decades, especially in the social sciences, I will use that excuse to discuss them first. Norm referencing, of course, implies that some quality or characteristic of something, usually an organism, is judged according to its relationship to that quality or characteristic represented by the mean or norm of the universe of those organisms. Thus, because intelligence is a relative characteristic, that is to say, it can only be expressed in relation to the intelligence of other organisms (usually other people), we must measure the intelligence (assuming it is definable and measureable) of either all the universe of the organisms or a suitably large and randomly selected sample of that universe, compute the mean of those measures, then compare our subject to that mean. Inasmuch as standard deviations provide a handy way of expressing variances from the mean, they are usually used in the social sciences, but there are, of course, other ways of expressing this difference. In any case, this is the essence of norm referenced measures, and as we shall try to point out, they serve some very useful purposes. Let us look at some of these.

Norm referenced measurement allows us to rank order people on a number of traits and characteristics, especially with respect to individuals who have particular abilities for success--as success is defined for the particular system involved. The Scholastic Aptitude Test, or as it is more commonly dubbed, the SAT, is typical of this kind of norm referenced measures, and provides an indicator of the relative academic abilities of school applicants compared to the universe of all those who took the test. As such, it serves a rank ordeting function, which, when judgments concerning chance of future academic success are required, is a very useful tool for making selections. There are those who will argue that such tests as the SAT do not really tap the inherent abilities for success in the real world, but that is beside the point--SAT is not designed for that purpose. Another arena where norm referenced measures have been widely used is in the recent flurry of implementing competency-based educational requirements in the public school systems. In this case many systems have been purchasing norm referenced tests and comparing their students' performance against the national norms. Outcomes exceeding the norms usually bring smiles of satisfaction to the parental faces, while lesser outcomes bring cries of anguish directed at the school authorities. Although it is nice to know that one's school system produces average scholastic test performance which exceeds the national averages, no parent really knows how well educated his child is becoming nor does he know how good or bad the national norm might be. As a matter of fact, the national norms on basic skill tests have been declining over the past several years, and, therefore, of course, performance at that norm may very well be no longer satisfactory if one really cares about performance on an absolute scale, or as compared with when mom and daddy went to school.

As we have said, norm referenced testing produces a rank order amongst those tested, and society demands such rank ordering in many of its segments, especially in education. Schools have traditionally provided this rank ordering through measures which are almost universally designed to provide a "spread" amongst the student's performances. In the customary setting of classroom, group oriented education the variable is the distribution of learning rate capability--the instruction is delivered to everyone at the same rate, and those who are rapid learners will show up well on the tests, while those who are slower will show up less well. The problem is that those under the lower half of the normal distribution curve of learning rate will not receive the same education as those under the upper half, and that is discrimination in the worst meaning of that term.

Norm referenced measures have a very real purpose in our world. When it is our need to describe the distribution of traits and characteristics of people in order to place individuals in their proper position in that distribution, there is no other way to go. But the danger is in assuming that all measures are for the purpose of placing people along some continuum as they relate to others in their performance or abilities--in some circumstances this may be useful, but there are many other circumstances where it is not, and in fact, results in the worst kind of judgments about people. This can occur in training situations if the constant is time to learn and the criterion is amount of learning taking place within that time constraint. Measures designed to produce a spread of test outcomes across a normal distribution under these circumstances are unsatisfactory and unfair to those who want to and can learn skills required for job proficiency, but need more time than the mean time to learn. Such a system is an unacceptable waste of human resources, and in the military especially, we must reject such practices.

Now, turning our attention to criterion referenced measures, let me attempt to both defend and castigate them. According to my dictionary, criterion is synonymous with standard--the two words mean the same thing.

Therefore, it is reasonable to define a criterion referenced measure as one which is referenced to a standard. By that definition, one may be excused for asking what measurement is not made against some standard--without being accused of playing with words to confuse the issues, one can surely not measure anything without a standard of some kind, be it a meter, a mile, a quart, a mean or a standard deviation. What, then, is so interesting about a criterion referenced measure, or, as we have observed, a standard referenced measure? Well, that is a reasonable question, but I suppose we are begging the issue--when educators speak of criterion-referenced-measures they really mean a measure which does not use the norm as its standard. So what we are talking about, then, are measures which have as their standard something other than the mean of the performance or behavior or characteristics or traits of the population of which the subject is a member. Inasmuch as such standards are not derived by formula, they are generally derived by fiat, and therein lies their weakness. "The student will, without assistance, orally recite the names of forty-nine of the fifty state capitols in the United States," is a criterion referenced measure, and we almost all would agree that it is a reasonably fair measure if taken in the context of a course in United States geography. But why has the designer chosen forty-nine of fifty as the criterion instead of fifty of fifty, or twenty of fifty, or any other combination? Instructional technologists who are really "with it," as the saying goes, would rather fight than use a norm referenced measure, but when you ask them to explain how they arrived at the criterion for their criterion referenced measure, they can easily become confused. Now I am not faulting criterion referenced measures, mind you. What I am trying to point out is that all that shines is not necessarily gold--even CRTs have their little flaws, and we should be aware of the pitfalls when we decide to go that route.

Many states have elected to establish competency standards for the award of high school diplomas, being somewhat sensitive to the criticisms that their graduates can neither spell, read nor perform simple arithmetic computations. Some of these states have elected to use as their standard the national norm for some grade, usually three or more grades below the twelfth, which causes me to ponder the rationale of a twelfth grade diploma judged on the basis of a ninth grade norm. Be that as it may, there are a couple of interesting points to be observed in this situation. Who decides what grade level to establish as evidence of high school education? The School Board? The Superintendent? The Parents? Somebody does, but of course the decision will please some and greatly displease others. What criterion should be chosen? Or should the school system design its own measures, either norm referenced to the local population or arbitratily standardized, such as my question about the fifty state capitols? The May 1978 issue of the Kappan Magazine (Kappan, 1978) is dedicated to the issue of minimum competence testing, and in it are four exemplars of school systems which have adopted minimum competency requirements and testing programs for their high school students. While I personally applaud the motives underlying these initiatives, the standards for achievement are as fuzzy as they have always been. In a recent issue of the local news paper in my town there was a set of test questions taken directly from the Florida minimal performance standards for twelfth graders. Here are a couple of samples, and I quote: "Mrs. Jones maps out a route from Atlanta to Key West. The distance is 780 miles. If Mrs. Smith drives this distance in 15 hours, what is her average speed." Another:

"Mr. Smith wants to buy a motorcycle for his son. He is interested in a good used cycle. Where could he find the information to help him?" My point in sharing with you these examples of so-called competency measures is that they are not standards of excellence, or even of mediocrity, but standards arbitrarily established to permit almost anyone who can read a primer to obtain a high school diploma. If this is the case, and I maintain that it is, then criterion referenced measures are as bad as, or worse than, norm-referenced measures in their lack of ability to discriminate the skilled from amongst the unskilled, which is the only excuse for a test measure I can think of. Now, on the other hand, those of us who have had some experience in technical training have found that we can come fairly close to establishing criterion referenced measures which are associated with skill levels determined as adequate for getting the job done satisfactorily. That sounds reasonable enough, does it not?

Well, let us look at a real world example. In the Naval Education and Training Command the curriculum for Apprentice Radiomen was recently completely revamped in accordance with established ISD procedures. One of the CRTs was something as follows: "Given a teletype machine and other required supplies, type a selection of textural material at the rate of 35 words per minute without error." Now the criterion of 35 words per minute was set by the Office of the Chief of Naval Operations, which in the Navy is tantamount to having come down on an engraved tablet from the mountain, and is plenty good enough for any criterion referenced measure developer. The only trouble was that in order to meet that criterion all the permissible time in the curriculum was taken with teletypewriter training, leaving no time for any other skills. Clearly the criterion was far too difficult for beginners in the Radioman rating, and was entirely unrealistic. But it took quite an effort to get it changed. The point to be made is that standards can easily be impractical as easily as they can be unchallenging. And in every case, if one really gets to the origins of any criterion, he will discover, underlying all else, a human judgment.

No discussion of measuring performance can overlook the growing importance of simulators for the task. If a simulator truly (that is, validly) requires the execution of skills which are required for the successful operation, or maintenance, of a machine, or system of some kind, then the measurement of the adequacy of the performance of those skills on the simulator can be accepted as a reliable measure of the adequacy of the performance when done on the real equipment or system. That axiom makes the utilization of simulators for this purpose most promising, and there is evidence of such application increasing. Among some of the more sophisticated examples of this technique is the simulator used by United Airlines to provide both training and proficiency checks to over 5,000 pilots yearly. Their system can determine and display quantitative scores, hard copy printouts of pilot performance, provide a simulator rerun of the pilot's actions, and display an optimum simulator run (General Physics Corporation Report, 1978). In this instance the values given performance are derived judgments supplied by forty flight instructors--as we have observed, standards stem from human decisions. In the Navy the submarine forces have developed, with the aid of the General Physics Corporation, a means by which audio tapes of enemy submarine noises can be fed to instruments manned by crew members who must "solve the tracking and fire control problem" from these data. Here the standard is a successful torpedo attack on the "enemy" submarine, and we are getting about as close to true criterion referenced measures as seems possible outside of the real world. One might adopt the old adage of "the proof of the pudding is in the eating therecf" as the definition of the perfect standard, if the performance gets the job done in the time needed to get it done, it passes criterion.

The purpose of this discussion has been, as I mentioned in the opening paragraph, to set before us some considerations about criterion and norm referenced measures. It has been my hope that we can agree to the observation that both have their respective purposes, each has its weaknesses and strengths, and must be carefully understood before acceptance as the panacea to any kind of measurements.

The Educational Testing Service Manual for Setting Standards on the Basic Skills Assessment Tests defines a standard as "an answer to the question, 'how much is enough?'" I think that is a very nice definition, and I commend it to your consideration. Of course the problem is the determination of that answer. Let me close with this quote, from the same publication. "It is extremely important to realize that all methods of setting standards depend on subjective judgment at some point in their application. There is no good way of setting standards just by plugging numbers into a formula."

Thank you.

REFERENCES

Naval Personnel Research and Development Center, Special Report 76TQ-14, July 1976

Naval Personnel Research and Development Center, Special Report 76TQ-15, July 1976

Educational Testing Service, Princeton, New Jersey, Manual for Setting Standards on the Basic Skills Assessment Tests, 1977

SQT PERFORMANCE STANDARDS

Gail T. Rayner

Florida State University
Center for Educational Technology
Tallahassee, Florida 32306

We have changed the name of the tests given to soldiers from MOS (military occupational speciality) tests to SQT (skill qualification test). With the name change there was to be a change in philosophy. Innovators were going to test performance not verbal information about performance. There was to be an annually increasing hands-on component and a decreasing written component. It hasn't turned out that way. We are testing very little "hands-on" and a lot of "written".

It is possible to test the tests. One could identify people who could perform a task using a hands-on test and one's who could not perform the test, and these same people could be given the written items. If there was a reasonable correlation between the results on the two tests we could say that they appeared to be testing the same thing. However, we do not test the tests that way. In addition, we set standards in some normative fashion.

Different types of standards can be set for measuring performance of military personnel in a variety of training and work settings and can fall into at least one of five categories:

1. Inherent standards
2. Empirical standards
3. Arbitrary standards
4. Part-task standards
5. Paper and pencil standards

This paper will be concerned with identifying the types of standards that are set in tests and the problems involved in the use of a particular type of standard. Examples of standards will be given in the discussion. Standards may be job standards, training standards described in the soldiers manual, performance testing standards or SQT standards

Inherent Standards

Inherent standards can be defined as those standards that are implied by the task itself, i.e., for a task involving equipment repair, the inherent standard would be that the peice of equipment would operate on completion of the repair task.

Inherent standards do not present problems for on-the-job performance but may cause difficulty when one sets approximate standards for training. For example, if the job task is "camouflage or conceal self and individual equipment," the job standards should be, "given a specific type of terrain, the person is concealed." However, the soldier's manual calls for having the shiny areas of skin shaded with dark paint stick and the shawdow areas with light paint stick; and clothing, equipment, and weapon outlines altered to blend with the backgorund within 15 minutes. The standard is 15 minutes, not that the individual or equipment is camouflaged, hidden, concealed, or difficult to see.

In the SQT, there are two illustrated items, one shows three faces with different patterns made by camouflage paint sticks, the other shows four helmets covered with various plant or man-made camouflage materials. The directions are: Given a situation, select the correct picture. The standard is to get both items correct or 100%. So we've gone from concealing oneself to identifying pictures of properly decorated faces and helmets.

Empirical Standards

Empirical standards are those standards based on how others who are deemed competent perform or on how engineers say the equipment can perform, i.e., from Army Material System Analysis Agency (AMSAA) data, estimates for the probability of hitting a moving target at 200 meters with an LAW is an empirical standard engineered in the LAW. Another empirical standard for the same weapon is the average probability of a hit based on firings of trained 11 Bravos. This second standard, based on experience, was aobut half that of the AMSAA standard.

For the job task "Load, reduce a stoppage, unload, and clear an M16A1 rifle" the standards are:

1. load and fire within five seconds
2. reduce a stoppage and then fire the next round within ten seconds
3. unload and clear within ten seconds

The times set are based on experience. Experienced 11 Bravos using an M16A1 can perform those tasks within the standards set. The soldiers manual task and standards are the same. However, the test items on the SQT are multiple choice questions on the most likely causes of stoppage and choosing the right next step in some described situations. The standard is 75%. So we have gone from the performance of some physical skill within a given time to a test on choosing the correct illustrated and written description of steps.

Arbitrary Standards

Arbitrary standards are just that--ones that are based on someone's judgement which may or may not be based on experience or data. Time limits are typical: Change a tire within five minutes. It could just as well be six or eight but five minutes is considered enough.

For a job task "Detect Enemy Mines," the job standard would be to find them without detonating them. In the soldier's manual , it says: Within 15 minutes, probe for and uncover buried mines without detonating. The arbitrariness for the job task of the 15 minutes standard is obvious.

The SQT consists of four true-false questions and three multiple choice questions on instruments for and methods of probing. The SQT standard is five out of seven. In this case we've gone from detecting without detonating a mine to verbal information on probing for mines.

Part Task Standards

There are many tasks which do not have specifiable standards for the whole task. For instance, one of the life saving measures is "Perform mouth to mouth resusitation and external heart massage". It means to perform it on someone who is not breathing and whose heart is not beating. A whole task standard could be that successful performance is measured by the victims recovery. But even correct performance cannot always resuscitate the victim. We cannot completely control the job situation. Our performance test standard applies only to correct procedures not to saving the victim.

To add to the problem one cannot apply these measures to a living non-victim; therefore, we have made simulators. There are dummies with soft faces and diaphragms. We can practice the skills on them and test on them. For years the standard has been "to the satisfaction of the instructor" plus a written test. Now there are dummies with built in feedback. If you perform the breathing correctly you will get a green light; if you place your hands on the chest incorrectly you will get a red light. For a real victim you are to continue until breathing and hear beat are restored or until help arrives. There isn't yet a simulator for that.

The SQT items are multiple choice information items. One item asks for correct proportions of chest compressions to breaths; the other asks how long you should continue performing first aid measures after you've gotten tired.

Paper and Pencil Standards

There are some tasks which are legitimate paper and pencil tasks; they are the paper and pencil tasks such as "Determine the grid coordinates of a point on a military map using the military grid reference system." The job task (if indeed this is a task not an element of another task) standards would depend on the degree of accuracy required for the situation. The soldiers manual standards are: Within 2 minutes determine the 6-digit grid coordinates for a point within 100 meters. The SQT has one multiple choice item which requires finding the 6-digit grid coordinate of a named building. In this one standards category--a paper and pencil test of a paper and pencil task--the items have face validity. In the other four discussed they do not.

Real World Standards

Given the SQT paper and pencil test and standards what do we really know about the soldiers ability to perform his job tasks? For the soldier, his promotion may depend on his passing the written test. For the Army, its defense capability is the sum of those soldiers' abilities to perform the tasks. An increase in the hands-on component and changes in the type of standards set for the SQT would more accurately reflect the soldiers' performance capabilities in the real-world settings they work in.

ON DEFINING JOB TASKS

Clay V. Brittain

Army Training Support Center
Fort Eustis, Virginia 23604

Last June when I submitted a title and abstract for this paper, I was vacationing on the beach in North Carolina. The paper seemed a much better idea from the leisured perspective of the oceanside than it does now. But having cast the die, let me begin by telling you where I am coming from in my concern about defining the term "task". For almost four years now my professional responsibilities have had to do with Army Skill Qualification Tests (SQT). The SQT Program has been described in a number of sessions at previous meetings of MTA. Many of you are familiar with it, especially those of you who work for the Army. But it is useful to quote a brief description of Skill Qualification Tests as characterized by Maier and Hirshfeld (1).

> "Skill Qualification Tests (SQT) have been developed to replace Military Occupational Specialty (MOS) proficiency tests as measures of ability to perform Army enlisted jobs. SQTs are performance-based, criterion-referenced measures of job proficiency, consisting of precisely defined tests of <u>tasks</u>, all of which are critical and necessary to performance of the job." (underlining mine).

As is indicated here, the concept of task is fundamental in Skill Qualification Testing. As criterion referenced measures of job proficiency, SQT's embody an approach to testing which grows out of the Instructional Systems Development (ISD) model. As you know, job/task analysis is the bedrock in the military application of the ISD model. Instructional objectives are specified in terms of job tasks to be performed. Thus, tasks are crucial in defining the domain of the SQT because tasks are crucial in the ISD model.

Before moving on from this point, let me comment about Soldier's Manuals. The Soldier's Manual is the companion document and major reference for the SQT. The Soldier's Manual lists and describes the tasks which have been identified as critical for soldiers in a given MOS. In other words, the Soldier's Manual tells soldiers what tasks they must be able to perform if they are to be fully competent in their MOS. The SQT tests a subset of tasks from the Soldier's Manual.

In perusing Soldier's Manuals, one is likely to be struck by the variation in the scope of task titles. The following examples perhaps are sufficient to illustrate this point. These are task titles which I selected from two or three Soldier's Manuals which I happened to leaf through.

Put on a protective mask

Administer antidote to a nerve agent casualty

Orient a map to the ground by map-terrain association

Measure ground difference by pacing

Monitor and evaluate training

Establish priorities for general maintenance

Conduct performance training

Counsel subordinates

The first four titles, of course, refer to quite specific activities which are narrow in scope. The last four, at least 3 of them, refer to broad classes or activities whose boundaries are at best ambiquous.

This type of situation poses serious problems for testing, especially for criterion-referenced testing, the very essence of which is the measurement of one's level of mastery of a performance domain. This requires that the domain be clearly specified. In a paper on content standard scores published before criterion-reference testing became fashionable, Ebel (2) commented on the importance of objectivity in test development. Ebel suggested that an objectively defined process of test development could be demonstrated if two test-developers, working independently of one another, came out with tests yielding similar results. In the present context, one could reasonably suppose that such objectivity might be rather easily achieved for a task such as "measure ground difference by pacing", but not for "monitor and evaluate training."

What is needed, as one informed observer has put it, is a better operational definition of task, for in the absence of a good operational definition, "task can stand for anything from a mission to turning a screw." This is a rather succinct statement of the problem which concerns me here.

As a person who is not extensively knowledgeable about the methodology or literature of job/task analysis, I began to wonder just what guidance might be found in the literature which would be relevant to the issue of defining job task. I wish that I could honestly report an exhaustive literature review and give you a definitive statement of what can be found in the literature. I cannot do that. But I can offer my own thoughts on what I have found bearing upon task definition. In a recent technical report on work sample testing, Guion (3) defines task as follows:

> A task is a preliminary statement, in rather broad terms, of a major activity, task, or responsibility of the job. It may be an appropriately formalized sentence such as "(Takes action)in (setting) when (action cue) occurs, using (tools, knowledge, or skill.

It seems to me that for some purposes, this level of precision in defining tasks probably is sufficient. In illustrating what I mean here, let me lift a statement from an article by McClelland (4) on the importance of competency testing. In reference to what he calls "criterion sampling," McClelland says:

> "If you want to test who will be a good policeman, go find out what a policeman does. Follow him around, make a list of his activities, and sample from that list in screening applicants."

Suppose that a bright and sensitive, but naive, observer followed policemen around with idea of describing their job activities. Such an observer would probably record and accumulate an enormous amount of data. In reducing the data to manageable scope, it would be necessary to summarize in terms of categories of recurring activity. These might be called tasks and one could find them very useful without agonizing about definitional precision. What one is after here is a more or less rough job description to serve as the framework for a more analytic description of job competencies. But it seems to me that one is in a much different ball game when engaged in criterion-referenced instruction or testing and the performance domain has been specified in terms of tasks. Precision of definition then does become important.

As I have indicated, the testing approach represented in SQT grows out of application of the ISD model to training. Task is fundamental in the ISD model. So in looking for guidance on how tasks are to be defined it seemed reasonable to turn to descriptions of the ISD model. TRADOC Pamphlet 350-30 is the Army document which sets forth the Interservice Procedures for Instructional Systems Development (ISD). In this document, several pages are devoted to the definition of duties, tasks, and task elements and to differentiating them from one another. It is of particular interest here to note what are said to be task characteristics.

1. A task statement is a statement of a highly specific action. The statement has a verb and object.

2. A task has a definite beginning and end.

3. Tasks are performed in relatively short periods of time, i.e., seconds, minutes, or hours, but rarely, if ever, days, weeks, months, or years. Although no definite time limit can be set, the longer the period of time between the beginning and the completion of the activity, the greater the probability that the activity is a generality or goal rather than a task.

4. Tasks must be observable in that by observing the performance of the job holder or the results of his efforts a definite determination can be made that the task has been performed.

5. A task must be measurable; that is, in the real world, a technically proficient individual can observe the performance of the task or the product produced by the task and be able to conclude that the task has or has not been properly performed.

6. Each task is independent of other actions. Each task statement must describe a finite and independent part of the job. Tasks are not components of a procedure. In the eyes of a job holder, a task is performed for its own sake in the job situation. A task is either performed or not performed by any one job holder. The job holder is never responsible for only part of a task. If he is responsible for only a part of a work activity that would otherwise be defined as a task, the part for which he is responsible is the task.

But it is noted in this treatment of the ISD model that most of these characteristics of tasks are also characteristics of task elements. It is only the last item (number 6) which differentiates a task from a task element. Before commeting on this, I wish to refer to a paper by Johnson and Richman (5) which was presented at the MTA Conference in 1975. This paper dealt quite lucidly with the problem of task definition. Johnson and Richman noted that with respect to task statements, the major problem has been with levels of specificity. Task statements on the one hand may be so broad as to be subject to various interpretations or, on the other hand, so narrow that they cannot be taken to refer to a task. They used examples to illustrate the consequences of task statements written at inappropriate levels of specificity. If you have not read this 1975 paper, I commend it to you. But in defining task, Johnson and Richman say "we insist that a task must be the smallest unit of job activity performed for its own sake in the eyes of a job incumbent in the job situation. I wish to comment about the statement that a task is performed for its own sake in the job situation. When I first came across this statement three or four years ago, I found it somewhat baffling. From an introductory course - I think in economics - many years ago I remember a similar statement that was offered as the distinction between work and play. That is, work is activity in pursuit of ends external to the activity itself. But play is activity engaged in for its own sake. Having observed the behavior of young animals - calves, puppies, kittens, children - this distinction intuitively made sense to me. But I am skeptical of the statement that a task is performed for its own sake. It seems to me that the very essence of task performance on the job is its instrumental nature, following the definition of instrumental offered by English and English (6). "Instrumental", they say, means "acting as a means to attain an end; characterizing that which is valued as a means to an end, in contrast with that which is valued for itself."

The statement that a task is performed for its own sake is puzzling to me, but it is also thought-provoking. Perhaps there is an implication here which is missed if the statement is taken too literally. In thinking about this, I remembered a study which a friend of mine did quite a number of years ago on the so-called Zeigarnik effect (7). This has to do with the persisting motivational effects of activities that are interrupted and thus not seen through to completion. The name comes from the name of a student of Kurt Lewin's who studied the phenomena at the University of Berlin in the late 1920's. Probably I am greatly oversimplifying, but Zeigarnik's studies ran something like this. Subjects were given a series of simple "tasks" to perform. They were allowed to complete some of them, but were interrupted and not allowed to complete others. Later when given a choice of activities in which to engage, including both the completed and incompleted tasks.the tendency was to choose the latter. It was as if failing to complete an activity in which one had become involved left one with a need for closure which could be gained only by resuming the task. The task had thus gained motivational properties of its own and in that sense was performed for its own sake. Parenthetically, such motivational properties of tasks I assume to be basic to Mager's (7) "Hey Dad" test. When ased "What are you doing now?" one is likely to respond in a way which points to activity sequences which give closure.

A major concern I have about the statement that a task is an activity which is performed for its own sake in the eyes of the job holder, is that it seems to imply that task definition is mainly subjective. That is, the difference between a task and a task element is in the eye of the doer. But I assume that no one would argue that this is, or can be, the case. In order to serve as a meaningful basis for training or testing, tasks must be amenable to objective specification. B. F. Skinner (7) talks about the importance in analyzing behavior of taking into account the "natural lines of fracture along which behavior and environment actually break." The idea seems applicable here. In the flow of human activity - including performance on the job - there are certain break points. A person completes one activity sequence and begins another. Thus in response to the "Hey Dad" test, it would make sense if one said "I am brushing my teeth". But "I am putting paste on my toothbrush" would seem odd. There is a break point here which seems inherent in the structure of the activity. It may be misleading to refer to break points for specifying job tasks as natural lines of fracture. They probably are largely reflections of how jobs are structured and responsibilities allocated among jobs. This I assume to be the import of a statement quoted earlier from the ISD model. "The job holder is never responsible for only part of a task. If he is responsible for only a part of a work activity that would otherwise be defined as a task, the part for which he is responsible is the task." It is a matter of what the job holder is tasked to do. Perhaps we could have a "Hey, Son" test. "What have you been told to do"?

Having brought up the idea of lines of fracture, what can be said about its application to task specifications? It certainly would not radically alter the fact that task specification is largely judgemental. One man's task is another man's task element. But the judgements of the analyst should reflect the purposes of the analysis. From the perspective of competency testing, the task should be of such scope as to lend itself to meaningful testing. It must be testable within the constraints of the test. For example, in the SQT program, the segment of the test (i.e. the subtest) which tests a task is the scorable unit. There are rather severe constraints on the size of scorable units. In the interest of testing which gives meaningful specific results, it seems necessary that the scope of the task be consistent with the limits of the test. What I am suggesting is a kind of corollary to the statement that in testing the test must be tailored to fit the task. I am suggesting here that the scope of the task must fit the test. The point is similar to one made in the Johnson and Richman paper cited above; namely, that the "trial-test-revise" logic employed in the development and validation of training materials is also applicable in developing task statements.

I conclude the paper with one other observation. Providing clear and useful guidelines on task definition is difficult largely because of the variety and complexity of human performance. Tasks take different forms. The description of the ISD model cited earlier differentiates unitary from multiple tasks. A unitary task is said to be one that is performed in the same way with the same inputs, e.g. disassemble on M16 rifle. Multiple tasks are of two types. One type is performed by basically following the same procedure, e.g. a task requiring multiplication of three-digit numbers by three-digit numbers. The other type of multiple task is one in which inputs vary, and the task is performed differently depending on the input; e.g., the military policeman's apprehension of a suspect. In dealing with the so-called unitary tasks, there probably is relatively little difficulty in developing task statements at appropriate levels of specificity. Defining multiple tasks of the first type, is somewhat more difficult. It may be problematic as to whether similar procedures are really basically the same.

For example, in arithmetic problems does it make a difference whether numbers are arranged vertically or horizontally? But it is with respect to multiple tasks of the second type, i.e. where situations might vary widely and call for extensively different responses that the definitional problem is so thorny. What might be called a task could emerge in an almost unlimited number of different versions. The problem here is more than that of determining an appropriate level of specificity. It is a matter of meaningfully grouping the concrete versions of the tasks to determine whether there is one task or many. It seems to me that in dealing with such multiple tasks, we are likely to move from an emphasis upon testing in terms of tasks to testing in terms of competencies (skills, knowledge, abilities).

1. Maier, Milton H. & Stephen H. Hirshfeld, Criterion-Referenced Job Proficiency Testing: A Large Scale Application, Research Report 1193, US Army, Research Institute For The Behavioral And Social Sciences. February 1978.

2. Ebel, Robert L. Content Standard Test Scores, Educational and Psychological Measurement Vol. XXII, No. 1, 1962 pp. 15-25.

3. Guion, Robert M. Principles of Work Sample Testing: III. Construction and Evaluation of Work Sample Tests, ARI Technical Report TR-79-A10, April 1979.

4. McClelland, David C. Testing For Competence Rather Than For "Intelligence", American Psychologist January 1973, pp 1-14.

5. Johnson, Robert N. and James N. Richman. Task Analysis: The Basis For Criterion Test and Curriculum Design; Proceedings of the 17th Annual MTA Conference, September 1975.

6. English, Horace B. and Ava Champney English. A Comprehensive Dictionary Of Psychological and Psychoanalytical Terms, Longmans, Green, and Co. Inc. 1961.

7. Green, D. R. Volunteering and The Recall of Interrupted Tasks, Journal of Abnormal and Social Psychology. 1963. Vol. 66, No. 4, 401-404.

8. Skinner, B. F. The Behavior Of Organisms. New York: Appleton-Century, 1938.

TASK ANALYSIS*

Major Robert B. Wiltshire, II

U.S. Army Signal Center and Fort Gordon

ABSTRACT

The presentation will address task analysis in contrast to a task list, provide a task analysis model which focuses on behavioral analyses and illustrate the application of task analysis to criterion referenced instruction with examples of test items and training modules. The task analysis model facilitates designing training products that are both effective and efficient in producing desired learning outcomes.

*This paper was presented but, due to its unavailability at the time of printing, only the abstract is reproduced here.

PREDICTION OF TANK GUNNERY USING JOB SAMPLES

Newell K. Eaton and Jimmy R. Johnson

US Army Research Institute for the Behavioral and Social Sciences
Ft Knox Field Unit, Ft Knox, KY 40121

INTRODUCTION

Prediction of tank gunnery performance has been the focus of a great deal of recent research (Greenstein and Hughes, 1977; Eaton, 1978; Eaton, Bessemer, and Kristiansen, 1979). In most of this research, paper-and-pencil aptitude/achievement tests have received primary focus as predictors. Job samples, which are short, critical components of the criterion performance requirement, have seldom been measured separately and used as potential predictors of overall performance levels.

The rationale for emphasizing paper-and-pencil tests is easy to understand. They are relatively inexpensive to produce and administer, easily scored, easily standardized, and very portable. Measurement of performance on job samples is much more difficult. Job samples usually require equipment on which to perform, and instrumentation for performance measurement. And such equipment/instrumentation is usually neither inexpensive nor easily portable.

Despite the research emphasis they have received, and their obvious practical advantages, paper-and-pencil tests have had only moderate success as predictors of tank gunnery performance. Test-performance zero order correlations greater than .35 are rare. Further, multiple regressions of test scores on gunnery performance infrequently exceed .50. Finally, predictor-performance relationships seldom cross-validate to new samples (Eaton, 1978; Eaton, Bessemer, and Kristiansen, 1979).

The limited research in which job samples have been evaluated offers more hope for improved prediction of gunnery performance. The few efforts that have been conducted have yielded numerous task-performance relationships in the .30 - .40 range (Eaton, 1978; Gobel, Baum, and Hagin, 1971). And job sample measures have yielded significant increments over paper-and-pencil tests alone in predicting training outcomes in several Air Force training programs (Hunter, Maurelli, and Thompson, 1977). Although no cross validation efforts have been made with job samples in the tank gunnery context, such efforts have been successfully conducted in other contexts.

There is also a logical reason to give job samples careful consideration as performance predictors. In the past, job performance was evaluated primarily with paper-and-pencil tests. Thus, such measures might be expected to share with paper and pencil predictor tests variance associated with cognitive test-taking skills. Today, there is a definite trend toward hands-on job performance evaluation. Indeed, all of our recent research efforts (Greenstein and Hughes, 1977; Eaton, 1978, Eaton, et. al., 1979) have emphasized hands-on performance criteria. Consequently, the predictive value of cognitive test-taking skills should be reduced, while the predictive role of psychomotor variables, such as job sample measures, should be enhanced.

Because of the limited success enjoyed by paper-and-pencil tests as gunnery performance predictors, the empirical and logical support for considering job samples as performance predictors, and the continuing need of the Army to predict who will perform well in tank gunnery, this research was designed to further explore the potential of job samples as gunnery predictors. The research was conducted in two phases. In Phase I a variety of job samples, including tracking, sensing, and round adjustment, were evaluated as potential predictors with tank loader/gunner trainees as research participants.

In Phase II, the most predictive job samples from Phase I were re-evaluated to determine whether the relationships observed in Phase I would be obtained with a second sample of tank loader/gunner participants. In addition, Phase II job sample performance was evaluated with a sample of research participants who were relatively unfamiliar with tank gunnery (tank driver trainees) to determine the effect of gunner/loader training on job sample performance. The information obtained was used to evaluate the extent to which job sample-gunnery performance relationships might be related to aptitude rather than achievement.

PHASE I

In this phase of the research, three job samples were chosen for evaluation. These represented three major requirements of tank gunnery performance: the requirement to operate tank controls in order to properly track a target (tracking); the requirement to sense the location of a fired round with respect to the target so as to be able to make a proper adjustment for a second round (sensing); and the requirement to change the point of aim, based on the location of a first round with respect to the target, to achieve a second round hit (round adjustment). Each of these requirements was tested with an appropriate simulator, yielding relatively objective performance measures. It was hypothesized that performance on the three job sample tasks would be positively related to tank gunnery performance. The three tasks seemed to capture, to a large degree, the requirements for successful tank gunnery.

METHOD

SAMPLE

Research participants were 47 men who were recent Armor gunner/loader training graduates at Ft Knox. Not all personnel were tested on all tasks due to time constraints. Of the 47 men, 26 were tested on the tracking task, 31 were tested on the sensing task, and 16 were tested on the round adjustment task. Of these, 10 men participated on both tracking and sensing, 10 participated on both tracking and round adjustment, 15 participated in both sensing and round adjustment, and 9 participated in all three tasks.

PROCEDURES AND VARIABLES

Crewmen participated in some or all of the job tasks described below. The three tasks were carried out in separate rooms, and scored by separate scorers.

Tracking Task. In the tracking task, gunner/loaders were given two geometrical designs, a diamond and a circle figure, to track on the Willey Burst-on Target (BOT) trainer. The Willey BOT Trainer is a device designed to simulate tank gunnery engagements. The equipment simulates the M60A1 tank fire control system and includes a set of gunner's hand controls which modulate the position of the gunner's sight reticle with respect to the target.

The task of each research participant was to track between the lines on each diamond and circle design a total of 20 times. Tracking direction (clockwise/counterclockwise) and design presentation (circle/diamond) were accomplished so that (1) each crewman started on a different design than the crewman preceding him, and (2) research participants alternated directions of tracking for each trial.

All crewmen received two scores for each of the 20 trials. The first score was the time required to track the entire design. The second score was an error score. This score was determined by the number of 1/2 second periods that the sight reticle on the Willey BOT trainer was outside the border line of the design. To determine the number of error periods for each trial, three scorers viewed the screen on the Willey BOT trainer and recorded the number of 1/2 second periods that the sight reticle was not between the double lines of the design. As with the time score, the error scores were the mean scores derived from the three scorers. Time and error scores for both figures were the means from the three scorers over trial 16-20 (the last five trials) for each crewman. Trials 16-20 were chosen because the asymptotic performance level was approached by trial 15.

Scores for time and error for both diamond and circle were raw scores. In addition, combined scores for circle and diamond were computed by converting each crewman's time and error scores (for each design separately) to z scores and adding the two scores together.

Sensing Task. On the sensing task, 31 men reported their sensing of a total of 25 simulated main gun rounds. Five target slides were prepared from 35mm slides taken of a tank gunnery range at Ft Knox, Kentucky. Each crewman was shown five simulated rounds for each of the five target slides.

The simulated main gun rounds were represented using a piece of red acetate covering a pinpoint hole made in an opaque slide. This slide, overlaid on a target slide, simulated the main gun round as it might appear going down range.

The simulation was accomplished by employing an ICONIX 6246 Tachistoscope, two Kodak carousel slide projectors and a standard screen. A target slide was projected on the screen for the duration of the five round simulated engagement. A second slide projector was controlled by the tachistoscope. It was set so that the simulated round would appear on the target slide for about 10 msec. The operator would announce "On the way" and manually start the tachistoscope for each of the 25 simulated rounds. After presentation of the simulated round, the research participant would plot his sensing of the round on a replication of the gunnery range target slide. The distance, in millimeters was measured between the true location of the simulated round and the location where the crewman had plotted the round. Each soldier's score was the mean of the 25 deviations.

Round Adjustment Task. The Fire Control Combat Simulator (FCCS) was used to measure the participant's round adjustment performance. The FCCS is a device designed to simulate tank gunnery engagements.

The task tested crewmen on their ability to apply standard burst-on-target (BOT) procedures while tracking a moving target. BOT is a technique of fire adjustment employed when a first round miss occurs. The program of instruction at Ft Knox indicated that all gunners were trained to note the position of the miss, relative to the sight reticle, and place that portion of the sight reticle on the center of mass of the target. Each gunner was forced to miss the first round, necessitating the use of BOT procedures to score a "hit" on the second round. Each crewman was given 6 practice engagements for familiarization, and 12 test engagements that were scored. Scores for the crewmen were the number of targets "hit" on the second round of the 12 two-round engagements.

Tank Gunnery Scores. Steel's Main Gun Tank Range at Ft Knox, Kentucky was used to collect tank gunnery performance measures. Stationary targets were 1.8m x 1.8m stationary plywood panels placed at ranges of 1000 meters and 1400 meters. The moving target was a 6.4m plywood flank tank target at approximately 700 meters, traveling at 5-10 km/h, and perpendicular to the line of fire.

Each crewman was required to fire a warm-up round at a 1200m panel, one two-round engagement at both the 1000 and 1400 meter targets, and two two-round engagements at the moving target. Scoring was achieved by two scorers observing each firing tank. One scorer used a tripod mounted 10x periscope while the other used a pair of 7x binoculars. The scorers knew when the gunner was about to fire each round, and the specific target at which he was to fire. Scores were recorded as a hit or miss, based on a consensus between the two scorers. Gunnery measures were number of moving target hits, number of 1st round hits, number of 2nd round hits, and overall scores, computed by multiplying the number of first round hits by 10 and adding that product to the number of second round hits multiplied by 5.

RESULTS

TANK GUNNERY RELATIONSHIPS

Tracking: Tank Gunnery Relationships. Correlations were computed between each of the tracking task measures (time, error, and combined scores, for both diamond and circle figures) and each of the tank gunnery measures (overall scores, moving target hits, 1st round hits and 2nd round hits). Although there were no significant relationships between circle tracking and tank gunnery performance, significant relationships were observed between diamond tracking and gunnery performance. Crewmen with fewer diamond errors had higher overall scores ($r = -.41$, $p < .05$), and more first round target hits ($r = -.50$, $p < .01$).

Sensing: Tank Gunnery Relationships. Correlations were computed between participants' sensing scores and their gunnery measures (overall score, moving target hits, 1st round and 2nd round hits). The relationship between sensing and both second round hits and overall scores approached significance (r = -.35 and -.34, p = .05 and .06, respectively). It is interesting to note that although relationships with the tracking task were statistically significant, and those with the sensing task approached significance, tho sensing and tracking tasks were not significantly correlated with one another.

Round Adjustment: Tank Gunnery Relationships. Analysis of the FCCS data and tank gunnery scores revealed no significant relationships. The ability to apply BOT procedures and hit a moving target on the FCCS device was not significantly related to the gunner's overall score, first and second round hits nor to the number of moving target hits.

Inter-Scorer Reliability on Tracking Error. The three scorers used in the tracking task provided high scorer reliability coefficients for diamond error (r's = .91 - .94) and moderate reliability for circle error (r's = .56 - .85).

DISCUSSION

The purpose of this research was to evaluate the relationship between performance on several job samples and tank gunnery performance. The results of the research revealed significant relationships between gunnery performance and both round sensing and tracking of the diamond figure. The fewer errors research participants made in sensing and tracking, the better they performed in tank gunnery.

The relationships with round sensing "make sense" and are readily interpretable. Round sensing, had a numerically higher correlation with second round hits than with first round hits. Such a relationship would be expected because round sensing, which involves knowing specifically where the gun just fired, and generally where it is firing with respect to the sight picture, provides specific information to the gunner on second rounds, but only general information to the gunner on first rounds of later engagements.

Relationships with diamond tracking error are not so easily interpretable. The tracking task should have measured a participant's ability to operate the turret/gun controls. While such ability should have been generally related to performance on all engagements, it should have been most strongly related to moving target hits. That, however, was not the case. Rather, diamond tracking was significantly related only to first round hits, and overall score. Interestingly, it was the error, rather than overall time measure, which accounted for these relationships. Thus it was not a gunner's overall speed on the tracking task, but instead his accuracy, that accounted for the variance in the gunnery performance.

Tracking the circle figure is much more difficult than the diamond. The circle requires constant changes in horizontal and vertical movement (or traverse and elevation), rather than a fixed degree of horizontal and vertical movement, with changes only at corners, as with the diamond. However, the gunners in this research were not required to perform such complex tracking in their main gun firing. The moving target ran slowly across the range in a constant direction. That might account for the failure to observe a relation with moving target hits.

It remains unclear why no predictor-criterion relationships were observed with the circle and overall gunnery performance, while such relationships did occur with the diamond. It is possible that the lower scorer reliability for the circle error may have contributed to the lack of significant correlations.

The failure to observe relationships between the round adjustment measure and gunnery performance is also not easily interpreted. The round adjustment measure should have evaluated both tracking and BOT ability. Yet despite the high fidelity of the FCCS controls, and the objective measurement potential it offered, no significant results with gunnery were obtained. In fact, the highest correlation, a -.21, was in the wrong direction, i.e., the more targets hit on the FCCS, the fewer main-gun second round hits. Apparently high-fidelity and sophisticated stimulus control and response measurement in a simulator are no guarantee of significant relationships with on the job performance.

PHASE II

Phase II of this research was designed to complement, and expand upon, the Phase I research. Overall, the results of Phase I research were favorable, and two job sample measures were related to gunnery performance (diamond error and sensing error). Several issues however, were not completely addressed.

First, because of the very small number of participants who performed all predictor tasks, the intercorrelations between tasks provided, at best, questionable estimates of population parameters. It was important to more fully explore these intercorrelations to determine the degree to which task performance accounts for unique gunnery variance. Tasks related to gunnery performance, but unrelated to one another, account for unique gunnery variance, and thus may be combined for improved gunnery prediction. Those tasks which are highly intercorrelated, on the other hand, are redundant, so a second measure adds little to prediction, once the first measure is known.

In addition, in the first phase of the research, many different measures of the job sample tasks were obtained, and many measures of main gun firing were evaluated, but the sample size was relatively small. Consequently, there was liberal opportunity for Type II errors in evaluating relationships between task measures and gunnery performance. Those tasks which did show apparently significant relationships with gunnery performance (tracking and sensing) needed to be re-evaluated to determine whether the task-gunnery relationships observed in Phase I were attributable to chance.

Finally, the task measures in Phase I were obtained from research participants who were completing training as tank gunner/loaders. Thus, the relationships observed may have been due to achievement rather than aptitude. Participants who received more help in training, had more prior experience, worked harder, etc., could have learned to be better gunner/loaders, and consequently perform better on both tank gunnery and the job sample tasks.

Such an hypothesis implies that, for some of the gunner/loader participants at least, training, aptitude, and/or motivation resulted in some level of achievement, and that the achievement was reflected by task performance. Consequently, this hypothesis can be evaluated by comparison of task performance of gunner/loader trainees, and other participants who are comparable in other ways, but are relatively untrained in gunnery.

It was the purpose of the Phase II research to address these issues by providing more stable measures of predictor task intercorrelations, re-evaluating and validating the significant task-performance relationships observed in Phase I, and determining whether task-performance relationships are more likely due to achievement or aptitude measurement.

METHOD

SAMPLE

Research participants included 24 men who recently graduated from Armor gunner/loader training at Ft Knox. In addition, the sample included 10 men who recently completed Armor driver training at Ft Knox. Unlike gunner/loaders, drivers were not given extensive gunnery training, and did not fire on main gun tank ranges at Ft Knox.

PROCEDURE AND VARIABLES

The tracking task and sensing task were conducted as in Phase I with only very minor changes. The round adjustment task was eliminated. Main gun firing was conducted as in Phase I. Only gunner/loaders completed main gun firing.

RESULTS

TASK GUNNERY RELATIONSHIPS

Tracking: Tank Gunnery Relationships. Correlations were computed between each of the tracking task measures and the tank gunnery measures. As in Phase I there were no significant relationships between circle tracking and tank gunnery performance, but significant relationships did exist between diamond tracking and tank gunnery performance. Gunner/loaders with fewer diamond errors had higher overall gunnery scores ($r = -.49$, $p < .02$), more moving target hits ($r = -.41$, $P < .05$), more 1st round hits ($r = -.43$, $p < .05$) and more 2nd round hits ($r = -.46$, $p < .05$).

Sensing: Tank Gunnery Relationships. Correlations were computed between the sensing score and the tank gunnery measures. A significant relationship was observed between number of errors on the sensing task and tank gunnery performance. Crewmen with fewer sensing errors had higher overall scores ($r = -.41$, $p < .05$). No significant relationships were observed between sensing score and the moving target, 1st round hits, or 2nd round hits, although the correlations with moving targets and 2nd round hits approached significance (r's = -.36 and -.35, $p < .10$). It is also worthy to note that the correlation between diamond error and sensing error was again non-significant (and in this case, zero).

Tracking and Sensing Combined: Tank Gunnery Relationships. Phase I research suggested the utility of both diamond error and sensing error as predictors of tank gunnery scores and indicated their intercorrelation was not significantly different from zero. Therefore, both error scores for Phase II subjects were combined as predictors of tank gunnery scores according to unit weighted model (as suggested by Einhorn and Hogarth, 1975; Schmidt, 1971; and Lawshe and Schucker, 1959). Unit weighted models have been suggested for small sample research because only the direction of the predictor-criterion relationship needs to be known: Beta weights are not computed, but instead are set arbitrarily at one. While there is somewhat less precision with this approach, there is also less opportunity for error in incorrect Beta-weight estimation.

In applying the unit-weighted model to the data, standardized scores were computed for tracking error, sensing error, and tank gunnery overall scores. Standardized error scores were then added together, and their sum correlated with standardized gunnery scores, yielding $r = .64$, $p < .01$ (the signs for the error score have been reversed to achieve a positive-going scale). Thus, the fewer errors a participant made in tracking and sensing the higher his predicted gunnery score, and the better his actual gunnery score. This relationship is shown in Figure 1.

In addition to computations with the unit weight model, standard multiple regression cross-validation techniques were employed. Beta weights for tracking and sensing error measures were computed from Phase I data and applied to standardized error scores in Phase II, yielding an $R = .65$, $p < .01$. The high similarity between unit model and multiple regression model results was due to the ratio of the computed Beta weights ($B_1 = -.64$, $B_2 = -.59$). Their ratio was .92, nearly 1.00. Consequently, in this case unit weight and multiple regression models provided nearly the same result.

Comparisons Between Gunner and Driver Performance. To evaluate the extent to which gunner's performance on the tracking and sensing tasks was a function of their prior gunnery training (achievement) in Armor gunner/loader training, comparisons of gunner's and drivers scores were made using t-tests.

DISCUSSION

The purpose of this phase of the research was to confirm and extend the findings of Phase 1. The results confirmed both of the significant relationships between tank gunnery scores and diamond and sensing error. In addition, both a unit weighted model and a standard multiple regression model, based on results from Phase I, provided a very good fit between Phase II diamond and sensing measures and gunnery performance. Approximately 35% of the variance on gunnery was accounted for by those two variables. The magnitude of this relationship was due to the fact that the two error measures proved to be uncorrelated, and thereby provided unique contributions to predictions of gunner's scores. In addition, the relationships between gunnery and job sample scores seem to be more likely due to aptitude rather than achievement measurement. This is because gunner/loaders who had considerable gunnery training, scored no better on the job sample tasks than drivers, who had relatively little gunnery training.

Overall, it appears that the development and empirical validation of an appropriate set of job samples gives promise of measures yielding reasonable large correlations with gunnery performance. Moreover, such relationships may be attributed, in large part, to aptitude rather than achievement measurement. Consequently, such techniques seem to have reasonable potential for use in assignment of personnel to appropriate training programs.

A more complete description of this research is contained in Eaton, N. K. and Johnson, J. R. Job Samples as Tank Gunnery Performance Predictors, Army Research Institute Working Paper FK 79-1, May 1979, available from the senior author.

REFERENCES

Eaton, N. K. Predicting tank gunnery performan-e (Research Memorandum 78-6. Alexandria, VA: US Army Research Institute, February 1978.

Eaton, N. K., Bessemer, D. W., and Kristainsen, D. M. Tank crew position assignment (DRAFT Technical Paper). Alexandria, VA: US Army Research Institute, February 1979.

Einhorn, H. J., and Hogarth, R. M. Unit weighting scheme for decision making. Organizational and Behavior and Human Performance, 1975, 13, 171-192.

Gobel, R. A., Baum, D. R., and Hagin, W. V. Using a ground trainer in a job sample approach to predicting pilot performance (AFHRL Technical Report TR-71-52). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, November 1971.

Greenstein, R. B., and Hughes, R. G. The development of discriminators for predicting success in armor crew position (Research Memorandum 77-27). Alexandria, VA: US Army Research Institute, December 1977.

Hunter, D. R., Maurelli, V. A., and Thompson, N. A. Validation of a psychomotor/perceptual test battery (AFHRL Technical Report TR-77-28). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, July 1977.

Lawshe, C. H., and Schucker, R. E. The relative efficiency of four test weighting methods in multiple regression. Educational and Psychological Measurement, 1959, 19, 103-114.

Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement, 1971, 31, 699-714.

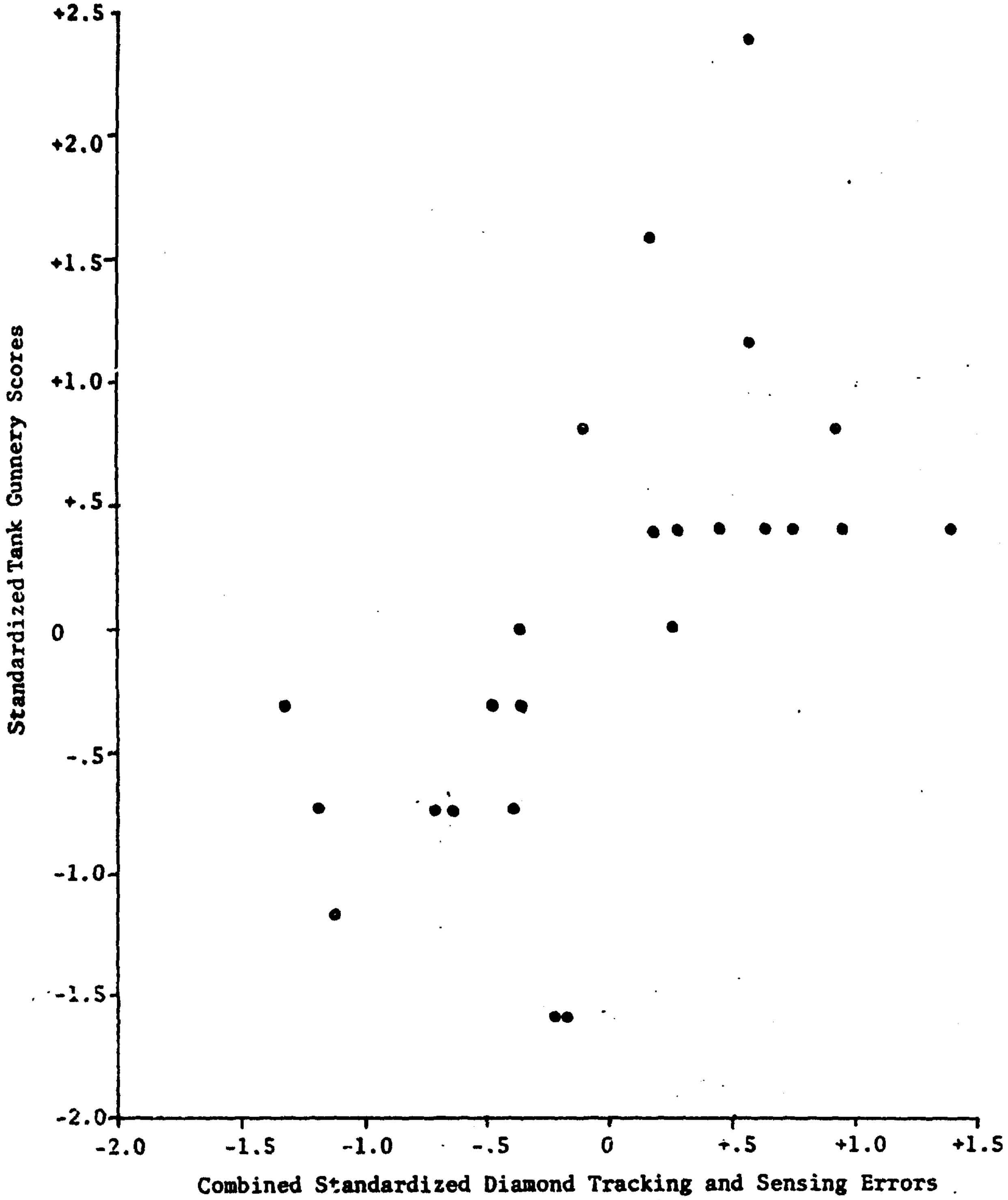


Figure 1. Prediction of Gunnery Performance from Combined Standardized Diamond Tracking and Sensing Errors Based on Unit Weight Model

IMPACT OF DESIGN AND SUPPORT PLANNING ON PERSONNEL REQUIREMENTS AND LCC*

H. Anthony Baran

Air Force Human Resources Laboratory

ABSTRACT

This paper describes a top-down approach to analyzing the impact of weapon system design and support planning on system support personnel requirements and Life Cycle Cost (LCC). The approach incorporates a modular system of both average value and simulation models, and a methodology for their application at successive stages of the Air Force systems acquisition process to effect increasingly accurate impact assessments. Applicable to both new and operational systems, it allows human resources requirements and LCC to be used more effectively as guideline considerations within both system design and modification processes.

The modeling system on which the approach is based provides for interactive operation of its modules and is user interactive to a degree which allows the operator to be an integral part of the system.

The top-down nature of the approach provides for analysis at: the system, subsystem or Line Replaceable Unit (LRU) level; various levels of design/planning definition; and various levels of detail. A key feature is that products are generated by an analytical rather than a parametric estimation process.

*This paper was presented but, due to its unavailability at the time of printing, only the abstract is reproduced here.

A Study of the Appropriateness of ANOVA as a Model for Comparing Differing Course Completion Times[1]

R. Eric Duncan

USAF Occupational Measurement Center
Randolph AFB, Texas 78148

A paper presented at the 21st Annual Conference of the
Military Test Association
October 1979

[1] The views expressed in this paper represent those of the authors and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

Abstract

Instructional evaluators and the organizations supporting them are seeking a practical and reliable statistical methodology to compare variable and constant course completion times. The present paper suggests that ANOVA is an appropriate model and presents three field examples to support this suggestion. The examples analyze the effects heterogeneity of variance has on the true alpha-level and the practical impact these effects have on decision-making. It was found that heterogeneity of variance had neither a practical nor statistically significant impact on the alpha-level suggesting that the possibility of Type I errors were not present above those at the level of null hypothesis rejection.

A Study of the Appropriateness of ANOVA as a Model for Comparing Differing Course Completion Times

R. Eric Duncan
USAF Occupational Measurement Center
Randolph AFB, Texas 78148

As instructional methodology moves toward a more individually-guided, self-paced mode, a proper and systematic approach for analyzing the differences between variable and constant course completion times is necessary. A major concern for decision makers is the confidence they have in a true difference between groups (i.e., how confident can they be that rejecting H_0 at .05 is reliable?). This concern originates from the effect that heterogenous variances (variable vs. constant times) and unequal cell sizes has on alpha. After a brief introduction to the assumptions of ANOVA, this paper will examine three situations that occured in a computer-assisted instruction (CAI) comparative field study, using real course completion data. (See Deignan and Duncan, 1978). The first situation was one in which variable and constant course completion times were compared. The situation was unique since it comparatively examined variance when the larger n had a smaller variance and the smaller n had a larger variance. The second situation, like the first, compared variable and constant course completion times, but in this situation the larger variance had the smaller n and the smaller variance had the larger n. The final situation compared two methods with variable course completion times but revealed heterogeneity in the direction as described in the second situation. This last situation also provides a look into how the search for aptitude treatment interactions (ATI) may have affected the significance of the obtained F. In all situations normality was tested before ANOVAs were conducted. Since Type I errors are of primary concern here, the effects of a loss of power on the F will not be discussed. Three procedures were used to examine the robustness of the F. The first was the classical ANOVA with no corrections for violated assumptions. The second method attempted to correct for unequal n's and variances by performing two separate t-tests, using formulas as specified by Edwards (1962) and Glass and Stanley (1970). The third method was an inhouse method derived by comparing differences from the grand mean (the constant value in those cases where variable and constant course completion times were compared).

Many research studies have been conducted to assess the effect of heterogeneous variances on the level of significance of the F-test (Scheffé, 1959; Lindquist, 1953; Boneau, 1960; Cochran, 1947; Godard and Lindquist, 1940; Horsnell, 1953; Welch, 1937) and cumulative conclusions are threefold (Glass and Stanley, 1970, p. 372):

1. When the sample sizes are equal, the effect of heterogeneous variances on the level of significance of the F-test is negligible.

2. When the sample sizes and variance- are unequal and fewer persons are sampled from the populations with larger variances, the probability of a type I error is greater than α. In other words, the effect of heterogeneous variances in this case is to shift the distribution of F-ratios to the right.

3. When the sample sizes and variances are unequal and greater numbers of persons are sampled from the populations with larger variances, the probability of a type I error is less than α. The effect of heterogeneous variances in this case is to shift the distribution of F-ratios to the left.

Since most experimental studies conducted in the field always desire, but seldom obtain, equal sample sizes, our concern should be focused on the second and third conclusions and their impact on data interpretations.

As shown in Table One, a one-way analysis of variance on time to complete instruction was conducted, and a significant main effect was found ($p<.002$). The sample sizes in the two cells of this design differed (93 vs 98), and the Bartlett's Box-F was significant, indicating heterogeneity of variance. Examination of the relationship between variances and n's revealed the smaller n had the larger variance and the larger n had the smaller variance. This situation suggests that the true alpha-level is something other than the obtained probability of .002. With this in mind, some might argue that a Type I error may have been committed in this test. In order to access the effect of heterogeneity of variance on the significance of the F-test and its true alpha-level, two correction factors were used to correct for unequal n's and unequal variances. The two formulas are shown in Table 2 with the appropriate calculations. As can be seen, the first formula (Edwards, 1962) reveals a significant difference between the groups as does the second correction factor (Glass and Stanley, 1970). These results indicate that the impact of heterogeneity of variance in the one-way analysis was not serious enough to effect the significance of the F-test. Indeed, the values obtained in the correction formulas when squared (3.11 and 3.19) did not substantially differ from the original F($F=10.21$).

(Insert Table One About Here)

(Insert Table 2 About Here)

There is another procedure which can be used to correct for heterogeneity of variance when a method using a constant value is compared with a method having varied values at different levels of a second variable (i.e., aptitude). By computing the differences between method 1 (the constant) and method 2 at all levels of the second variable, an overall difference between methods can be obtained. When this overall difference ($\bar{X}...$) is divided by the standard error of the differences ($sd/\sqrt{n-1}$) (a variance estimate which corrects for unequal n's and unequal variances) the resulting t reveals the difference between the method with the constant score and the method with varying scores. Table 3 illustrates this method for CAI time to complete instruction when compared with constant lecture

(Insert Table 3 About Here)

times. Even though the conservativeness of this test and its associated reduced significance level due to markedly fewer degrees of freedom ($t=3.10$, $p < .01$) was not as great as that obtained in the F-test ($F=10.21$, $p < .002$), the heterogeniety of variance did not affect the practical decision to reject the null hypothesis. This procedure offers two by-products which (1) allow us to examine the aptitude main effect as seen in Table 4 and (2) enables us to examine the significance of differences between CAI times and the constant

(Insert Table 4 About Here)

lecture times for each aptitude level as seen in Table 3.

The identical procedures employed in Medical Lab were also used to examine the Dental course time data. The major difference in the present case was the relationship of sample sizes to variances. In the Medical Lab course, the larger n had a smaller variance and the small n had a larger variance, which warranted the conclusion of an increased possibility of Type I errors. Whereas, in the Dental course the larger n had the larger variance (CAI) and the smaller n had the smaller variance, suggesting that there was a smaller chance for a Type I error.

The one-way analysis of variance, as seen in Table 5, reveals a significant Bartlett's Box -F which indicated heterogeneity of variance. The direction of that heterogeneity is explained above. The same formulas used to correct for unequal n's and unequal variances in the Medical Lab course were used in the Dental course. Table 6 shows these correction formulas and the probability levels associated with the corresponding

(Insert Table 5 About Here)

(Insert Table 6 About Here)

obtained t's. When these t-values were squared ($t^2 = F$), values for the Glass and Stanley correction approximated (43.43), and the Edwards formula doubled (84.64) the original F-value ($F = 41.98$, $p < .001$). These results support the conclusion that, with variances and n's aligned proportionately, fewer Type I errors would occur.

The second procedure for correcting for unequal n's and unequal variances, as mentioned previously, was employed in the Dental Course. Since fewer degrees of freedom are used in obtaining the significance level of the t determined by this test than in the ANOVA, we are reducing the chance of finding a difference. The test, however, revealed significance far in excess of the .001 level. Thus, this procedure again verifies the results of the original ANOVA, and in doing so, rejects the idea that heterogeneity of variance in this case and those similar to it makes the results of the ANOVA uninterpretable. As by-products of this test, we were able to verify significant differences at each aptitude level as seen in Table 7 and to show a significant main effect for aptitude ($F = 8.88$, $p < .0001$) as shown in Table 8.

(Insert Table 7 About Here)

(Insert Table 8 About Here)

A final situation differs from those previously presented, in that, both methods under consideration have unequal variances, as indicated by a significant Bartlett's Box-F, and are continuous. The direction of the heterogeneity in this case is one in which the larger variance has the smaller n and the smaller variance has the larger n. Given these constraints, one statistical approach which uses the correction formulas previously mentioned appears warranted.

The results of the questioned one-way analysis of variance appear in Table 9. As shown in Table 9, there was a significant main effect for treatment, ($F=4.32$, $p=.044$). Due to the relationship of n's and variances, as mentioned above, there is a possibility of a Type I error. By using the correction formulas previously used, this possibility can be assessed. Table 10 presents the results of these correction formulas which indicate that the heterogeneity of variances does not practically impact the decision to reject the null hypothesis at the .05 level. However, the results of the correction formulas closely approached, but exceeded the critical $_{.975}t_{,\infty}$ of 1.96.

(Insert Table 9 About here)

(Insert Table 10 About Here)

Since ATI's were of interest to the experimenters, a 2 X 3 analysis of variance was conducted and the results are seen in Table 11. Twenty-eight students lacked precourse assessment data and therefore their criterion scores were not used in the second analysis. The main effect for treatment was not significant at the conventional level ($p < .05$), but approached significance ($F = 3.31$, $p = .07$). There are two possible interpretations

(Insert Table 11 About Here)

of these differing results: (1) Since the corrected t-values closely approximated the critical t-value, measurement error could be causing the conventionally significant difference; or (2) the difference revealed by the 2x3 ANOVA treatment main effect may be a true reflection of the population differences (i.e., no differences).

To recap, three separate experimental situations were examined to determine the effect heterogeneity of variance had on true alpha-levels. The first analysis examined provided for a smaller n with a larger variance and a larger n with a smaller variance when variable and constant times to completion were compared. The possibility of committing a Type I error was not supported in this situation. The second and third analyses also did not significantly nor practically effect the true alpha-levels of their respective ANOVAs.

Even though classical theory invalidates the use of ANOVA when its assumptions have been violated, practical application of the results obtained here provides the instructional evaluator a method for the comparative evaluation of variable and constant course completion times in which the evaluator can be made: (1) It is necessary to strictly control data collection to ensure minimal data loss and non-random subject elimination, (2) to permit flexibility in experimental recommendations which would provide decision-makers with practical choices not invalidated by violations of classical statistical theory, and (3) strict adherence to classically accepted alphas may not provide an evaluator with the information necessary to make a practically significant decision.

BIBLIOGRAPHY

Bartlett, M. S. "The effect of non-normality on the t distribution." Proceedings of the Cambridge Philosphical Society, 31 (1935), 223-31.

Boneau, C. A. "The effects of violations of assumptions underlying the t-test." Psychological Bulletin, 57 (1960), 49-64.

Cochran, William G. "Some consequences when the assumptions for the analysis of variance are not satisfied." Biometrics, 3 (1947, 22-38.

Deignan, Gerard M. and Robert E. Duncan. "CAI in Three Medical Training Courses: It was Effective!", Behavioral Research Methods and Instrumentation, 1978, Vol. 10 (2), 228-230.

Edwards, A. L. Statistical Methods for the Behavioral Sciences. New York: Holt, Rinehart and Winston, 1962.

Glass, Gene V. and Julian C. Stanley. Statistical Methods in Education and Psychology. New Jersey: Prentice-Hall, 1970.

Godard, R. H. and E. F. Lindquist. "An empirical study of the effect of heterogeneous within-groups variance upon certain F-tests of significance in analysis of variance." Psychometrika, 5 (1940), 263-74.

Horsnell, G. "The effect of unequal group variances on the F-test for the homogeneity of group means." Biometrika, 40 (1953), 128-36.

Lindquist, E. F. Statistical Analysis in Educational Research. New York: Houghton Mifflin, 1940.

Scheffe, Henry. The Analysis of Variance. New York: John Wiley, 1959.

Welch, B. L. "The significance of the difference between two means when the population variances are unequal." Biometrika, 29 (1937), 350-62.

TABLE 1

One-Way Analysis of Variance Before Correction for Unequal N's and Unequal Variances for Medical Lab Time to Complete Instruction

SOURCE	df	SUM OF SQUARES	MEAN SQUARE	F
CAI vs Lecture	1	240542.80	240542.80	10.21***
Within	189	4452800.30	23559.79	
Total	190	4693343.10		

***p < .002

TABLE 2

T-Test Results after Correction for Unequal N's and Unequal Variances for Medical Lab Time to Complete Instruction

EDWARDS

$t= \bar{X}_1 - \bar{X}_2 / \sqrt{s_1^2/n_1 + s_2^2/n_2}$

$t= 540 - 469/ \sqrt{0/98 + 48400/93}$

$t= 71/ \sqrt{520}$

$t= 71/ 22.81$

$t=3.11 \qquad p < .01$

GLASS AND STANLEY

$t= \bar{X}_1 - \bar{X}_2/ \sqrt{(n_1-1)s^2 + (n_2-1)s^2 / n_1 + n_2 - 2 \quad (1/n_1 + 1/n_2)}$

$t= 71/ \sqrt{23559.79\ (.02096)}$

$t= 71/22.22$

$t= 3.20 \quad p < .01$

TABLE 3

The Use of the Grand Mean of Difference Scores and the Standard Error of the Difference Scores to Test the Difference between Lecture Time to Complete Instruction at All Aptitude Levels (a constant of 540) and CAI Time to Complete Instruction at All Aptitude Levels in the Medical Lab Course.

$\bar{X}_{..} = 76.000$
Se = 24.52
n = 86

Low aptitude difference = 32.05
Middle aptitude difference = 56.64
High aptitude difference = 192.31

$t = \bar{X}_{..}/Se$
$t = 76/24.52$
$t = 3.10^{**}$

t (Low apt) = 1.31
t (Mid apt) = 2.31*
t (High apt) = 7.84***

$^{*}p < .05$
$^{**}p < .01$
$^{***}p < .001$

TABLE 4

Analysis of Variance of Difference Scores (a constant 540 minutes - CAI block time) for CAI Time to Complete Instruction at Various Aptitude Levels in the Medical Lab Course.

SOURCE	df	SUM OF SQUARES	MEAN SQUARE	F
Aptitude	2	611490.54	305745.27	6.70***
Within	83	3784563.47	45597.15	
Total	85	4396054		

$^{***}p < .001$

TABLE 5

One-way Analysis of Variance of Time to Complete Instruction Within the Dental Course Before Correction for Unequal N's and Unequal Variances

SOURCE	df	SUM OF SQUARES	MEAN SQUARES	F
CAI vs Lecture	1	255931.13	255931.13	41.98***
Within	151	920526.87	6096.20	
Total	152	1176458.00		

***p < .001

TABLE 6

T-Test Results after Correction for Unequal N's and Unequal Variances for Dental Course Time to Complete Instruction

Edwards

$t = \overline{X}_1 - \overline{X}_2 / \sqrt{s_1^2/n_1 + s_2^2/n_2}$

$t = 87/\sqrt{0/52 + 9025/101}$

$t = 87 \ / \ 9.45$

$t = 9.20$ $p < .001$

Glass and Stanley

$t = \overline{X}_1 - \overline{X}_2 / \sqrt{(n_1-1)s_1^2 + (n_2-1)s_2^2 / n_1 + n_2 - 2 \ (1/n_1 + 1/n_2)}$

$t = 87 / \sqrt{(51) \ 0 + (100) \ 9025 \ / \ 151}$

$t = 87 / \sqrt{174}$

$t = 87 \ / \ 13.19$

$t = 6.59$ $p < .001$

TABLE 7

The Use of the Grand Mean of Difference Scores and the Standard Error of the Difference Scores to Test the Difference between Lecture Time to Complete Instruction (a constant of 540) and CAI Time to Complete Instruction at All Aptitude Level's for the Dental Course

$\bar{X}$ ... = 86.70	Low Aptitude Difference = 43.17
Se = 11.32	Middle Aptitude Difference = 83.57
	High Aptitude Difference = 153.05
	t (Low Apt) = 3.81***
t = $\bar{X}$../Se	t (Mid Apt) = 7.38***
t = 86.70/11.32	t (High Apt) = 13.52***
t = 7.66***	

***p < .001

TABLE 8

Analysis of Variance of Difference Scores (a constant 540 minutes - CAI block time) for CAI Time to Complete Instruction at Various Aptitude Levels in the Dental Course

Source	df	Sum of Square	Mean Square	F
Aptitude	2	149570.73	74785.37	8.88***
Within	76	640087.98	8422.21	
Total	78	789658.71		

***p < .0001

TABLE 9

One-Way Analysis of Variance of Achievement Scores within the Radiology Course Before Correction for Unequal N's and Unequal Variances

Source	df	Sum of Squares	Mean Square	F
CAI vs PIT	1	61.56	61.56	4.32*
Within	184	2620.57	14.24	
Total	185	2682.13		

$^{*}p < .05$

TABLE 10

T-Test Results after Correction for Unequal N's and Unequal Variances for Achievement in the Radiology Course

Edwards

$$t = \overline{X}_1 - \overline{X}_2 / \sqrt{s_1^2 / n_1 + s_2^2 / n_2}$$

$$t = 2.77 / \sqrt{(10.62)^2 / 89 + (8.14)^2 / 97}$$

$$t = 2.77 / \sqrt{1.95}$$

$$t = 2.77 / 1.397$$

$$t = 1.98 \qquad p < .05$$

Glass and Stanley

$$t = \overline{X}_1 - \overline{X}_2 / \sqrt{(n_1-1)s_1^2 + (n_2-1)s_2^2 / n_1 + n_2 - 2 \; (1/n_1 + 1/n_2)}$$

$$t = 2.77 \; / \; \sqrt{(88.51) \; (.0215)}$$

$$t = 2.77 \; / \; 1.38$$

TABLE 11

Analysis of Variance of Achievement Scores for PIT - CAI Treatments and Aptitude Level Conditions in the Radiology Course

Source	df	Sum of Squares	Mean Square	F
Main Effects				
CAI vs PIT (A)	1	42.700	42.700	3.31
Aptitude (B)	2	161.716	80.858	6.26**
AXB Interactions	2	238.107	119.054	9.22***
Within	149	1922.588	12.903	
Total	154	2371.187		

**p < .01
***p < .001

CLUSTER ANALYSIS VERSUS FACTOR ANALYSIS IN DEFINING JOB GROUPS

Jay A. Gandy
U.S. Office of Personnel Management
Washington, D.C. 20415

INTRODUCTION

Effective personnel mangement requires the classification or grouping of similar jobs for a number of purposes. A variety of methods have been developed for grouping individuals or other entities which are highly similar on specified attributes of interest. Two general categories of methods which have proven useful are clustering techniques and cluster-oriented factor analytic techniques. According to Overall and Klett (1972), "Although the methods used to accomplish classification grouping vary, they have two major steps in common: (1) computing quantitative indices of multivariate similarity between all pairs of individuals or objects and (2) analyzing similarity indices to identify homogeneous subgroups" (p. 182).

This paper reports on the comparison of factor analysis and cluster analysis in the context of a problem faced by the Office of Personnel Management in validity generalization. We needed to determine whether the clerical tests used for entry level hiring for a number of clerical jobs in the Federal Government can appropriately be used also by State and local governments. The Intergovernmental Personnel Act encourages governmental jurisdictions at various levels to enter cooperative recruiting and examining agreements which, among other purposes, serve to more efficiently use resources for government hiring activities and provide greater convenience to the applicant public. Such an agreement operates in Utah where the Federal Government (OPM) participates with the State of Utah and several county and local jurisdictions in a cooperative agreement under which applicants can go to one location, take one test, and potentially qualify for clerical jobs in multiple jurisdictions.

A need existed to ensure that jobs of participating governments could be filled appropriately using the Federal test, i.e., that the validity of the test could be generalized to additional jobs. Under the Uniform Guidelines on Employee Selection procedures (U.S. Equal Employment Opportunity Commission et. al., 1978) incumbents in jobs to which validity is being generalized must "...perform substantially the same major work behaviors, as shown by appropriate job analysis..." (Section 7B2).

Research highly relevant to this problem has been done recently by Pearlman and Schmidt (Note 1) based on the general solution to the problem of validity generalization developed by Schmidt and Hunter (1977). They analyzed validity data from published and unpublished clerical validation studies including over 2700 validity coefficients. The jobs involved were categorized according to clerical job families of the Dictionary of Occupational Titles (D.O.T.), (U.S. Dept. of Labor, 1977). Their research showed that when sources of artifactual variance (e.g. sampling error and differences among studies in test reliability, criterion reliability, and range restriction) are taken into consideration, a number of test types are clearly valid for a number of job types.

One objective of the present study was to determine whether D.O.T. job family categories could be derived empirically -- as opposed to judgmentally -- from task ratings obtained from incumbents, where the incumbents indicate which tasks they perform and estimate the relative amount of time they spend on each task. Such a procedure could strengthen the basis for conclusions on validity generalization by confirming the relationship between each job and the job families for which validity has been previously demonstrated.

There is evidence that Ward's, (1963) hierarchical grouping technique used in the CODAP programs developed by Christal (Note 2) and associates is the best of a number of available clustering procedures. Blashfield (1976) systematically compared four of the most popular agglomerative hierarchical clustering techniques and found Wards's minimum variance procedure to consistently yield the most accurate grouping. (The criterion groups were predetermined on a mathematical basis.) Blashfield warns, however, "...that different methods can yield very different solutions and that users should be careful to skeptically test the classifications generated by cluster analysis methods" (p. 377).

Overall and Klett (1972, p. 201) recommend factor analysis as an additional, or alternative, method for grouping individuals or things. They point out that cluster analysis methods may be subject to problems of stability and replicability since the starting point for each empirical group or cluster depends upon only two or three profiles out of the total sample. Factor analysis methods, on the other hand, take into account relationships among all individuals in the sample and may be more likely to result in more consistent and reproducible groups.

Factor analysis for identifying clusters of individuals or jobs differs from the more usual factor analysis procedures in the organization of the data (correlation matrix) which is used as input. In the more usual procedure each pair of variables is correlated across people. In contrast, for clustering purposes, each person, or job is correlated with each other person, or job, across variables. In other words, correlations are obtained between each pair of profiles which describe the individuals, or jobs. The basic data is the same for either factor analysis procedure, but through transposing rows and columns a different correlation matrix results. Correlations between profiles are usually called Q-type correlations, as opposed to the more frequently used R-type correlations.

METHOD

Instrument Development

A CODAP survey task inventory was developed similar to those used routinely in military contexts. Attention was paid in the development process to ensure: (a) comprehensive coverage of work performed, (b) clarity of task statements and instructions, (c) consistency in level of task specificity, and (d) meaningful organization of tasks under duties.

The instrument was pretested, reviewed by all participating jurisdictions, and revised as needed. The final task inventory contained about 300 clerical tasks grouped under 13 duties plus one "miscellaneous tasks and duties" category.

Data Collection

The data treated in this anlaysis included 1357 incumbents from Federal agencies in Utah, the State Government, and 19 local jurisdictions. Incumbents were included from 175 job classes. Jobs with the same title but covering multiple pay levels were frequently combined as a single class.

The sampling plan was keyed to the population in the job class with larger percentages sampled for smaller classes. Overall, approximately 39% of the population was sampled. The number sampled per class ranged from one to 158. Of the 175 job classes 74 were represented by a single incumbent. These were almost always single incumbent jobs in small local jurisdictions which essentially had no classification system.

Preliminary Analysis Relating to Quality of Instrument and Data

Various checks were made on the quality of the data and instrument. Incumbents did not tend to inflate their ratings or number of tasks performed, and they generally believed the inventory coverage, scope, clarity, etc. was good.

In order to evaluate the grouping of tasks under duties, a regular R-type factor analysis was performed on the intercorrelations of the 14 duties across the 175 job classes. The data consisted of the average percentage of time spent by each job class on each of the 14 duties. (It is noted that most incumbents in most job classes reported time spent in almost all duties; thus there were a fairly small number of zeroes for duty areas.) A factor analysis using ones in the diagonal and varimax rotation resulted in a clear simple structure with 13 factor accounting for 100 percent of the variance. Twelve duties loaded between .92 and .99 on each of 12 factors. The thirteenth factor was bipolar with the duty "bookkeeping and financial duties" loading .60 and "interacting with people" loading -.87. The percentage of variance accounted for by each factor ranged from 7.1 to 8.8 percent. These results were interpreted as indicating that the duty structure used in the task inventory reflected essentially independent dimensions and should be used without modification in further analysis.

Cluster Analysis and Factor Analysis

As indicated previously, the major analyses were cluster analysis and factor analysis. It should be mentioned that the cluster analysis included 168 additional incumbents who were not included in the factor analysis. An objective was to include in the study only those incumbents in clerical jobs from entry level to that level which is reached by the majority of incumbents within approximately five years. Incumbents deleted from further analysis were identified on the basis of background variables as clearly at a level higher than this. Their removal appears to have had no effect on comparisons of interest in this study.

For the factor analysis composite data for each class was used, giving the average percentage of time spent on each duty by each job class. This was done to rule out the effect of the grossly different $\underline{n}$ sizes of the job classes. Otherwise, the results of the factor analysis would be biased toward characteristics of the large

Table 1

Clusters of Clerical Positions Selected for Analysis

Group No.	KPATH Nos.	n	Overlap Between	Overlap Within	Aggregative Mean Overlap Between	Aggregative Mean Overlap Within
273	1 - 648	648	34.7%	42.4%	34.7%	42.4%
263	649-731	83	34.4	38.7	34.4	40.7
274	732-784	53	34.8	39.5	34.8	39.8
(various)	785-788	4			32.6	39.7
300	789-799	11	35.6	46.5		
(various)	800-803	4			30.8	39.4
335	804-811	8	37.0	47.8		
(various)	812-816	5			28.1	37.6
181	817-879	63	30.9	35.2		
(various)	880-890	11			26.6	37.4
169	891-932	42	30.3	36.0		
(various)	933-941	9			26.4	36.2
108	942-956	15	26.5	34.6	26.2	35.9
135	957-979	23	28.5	37.1	25.9	35.4
(various)	980-985	6			24.1	35.3
94	986-1091	106	25.6	31.4	23.5	33.2
(various)	1092-1095	4			22.5	33.1
91	1096-1159	64	25.4	30.5		
(various)	1160-1161	2			22.0	31.9
84	1162-1170	9	24.4	39.4	20.2	31.7
68	1171-1195	25	22.3	32.5	19.0	31.2
(various)	1196-1200	5			17.9	31.1
47	1201-1230	30	19.5	24.6	17.7	30.5
57	1231-1246	16	21.0	30.1	17.3	30.1
48	1247-1262	16	19.5	28.1	16.4	29.8
32	1263-1280	18	16.9	25.9	15.9	29.4
(various)	1281-1290	10			13.4	29.2
35	1291-1313	23	17.5	28.0	13.1	28.6
37	1314-1326	13	17.5	33.5	11.9	28.3
66	1327-1441	115	22.1	28.6		
(various)	1442-1450	9				
45	1451-1458	8	19.3	35.1		
(various)	1459-1462	4			11.0	25.3
28	1463-1481	19	15.9	32.6	9.6	24.9
(various)	1482-1525	44				24.0

classes. Additionally, the purpose of the study was to make discrete decisions about each job class but not about the possible misclassifications of individuals. Thus, the duty profiles of the 175 job classes were intercorrelated as input to factor analysis.

RESULTS

Cluster Analysis

The cluster diagram produced by CODAP began with 71 clusters. The within-group overlap of the groups ranged fairly closely around 50%. At the completion of clustering, with all incumbents in a single cluster, within-group overlap was 24%. A problem was evident with respect to defining a point in the hierarchical clustering at which to select clusters for further analysis. There were no major gaps at which between-group and with-in group overlap changed appreciably with subsequent clustering. The changes were very gradual.

A point was selected at which there were 22 clusters containing 92% of the sample, with the remaining incumbents lying essentially between clusters. Table 1 captures the clustering process at the point chosen for further analysis. In most cases the next step in the clustering process was for each group (or outliers, labelled "various" in the table) to merge, in the sequence listed, with the largest cluster.

The largest cluster, listed first in the table, contained 648 members which was 42% of the sample. With 42.4% within-group overlap, this cluster was also more homogeneous than most of the other, smaller, clusters. Two other clusters had slightly over 100 members, and the remaining 19 clusters ranged between eight and 83 members each, with moderate within-group overlap ranging between 25 and 48 percent.

A key question of interest was whether incumbents of the same job class joined the same clusters. Since the question is meaningful only for job classess containing multiple incumbents, the composition of the clusters was reviewed with respect to the 101 job classes which had two or more incumbents.

In only about half the cases (51 of 101 job classes) did the majority of incumbents of a job class belong to the same cluster. About two-thirds (33) of these job classes grouped under the largest cluster.

A similar pattern existed across most jurisdictions. The correlation between total number of job classes (with multiple incumbents) of a jurisdiction and the number of job classes of that jurisdiction in which the majority of incumbents joined the same cluster was .93 ($p < .01$) across the 17 jurisdictions having multiple-incumbent job classes. In other words, it was typical of these jurisdictions, large and small, that about half of their clerical classes cluster well (based on task level data) and about half do not.

Factor Analysis

As mentioned previously, the factor analysis procedure for clustering used a correlation matrix (175 x 175) composed of correlations between the profiles of each job class (mean time spent on each of 14 duties). Factor analysis was performed using ones in the diagonal, resulting in ten factors accounting for 99% of the variance.

A varimax rotation resulted in the clearest simple structure. The first four factors accounted for 85% of the variance, indicating a small number of primary job types in the data. Nine factors were readily interpreted due to the close relationship between factor loadings and the time spent measures on the task inventory. Every job class loaded .50 or higher on one of eight factors.

The identified factors and number of job classes loading highest on each are shown in Table 2. By far the largest number of job classes (46%) grouped in Factor I filing, sorting, routing, and related activities.

A relatively straightforward relationship exists between the primary factors and D.O.T. clerical job families. This linkage and the number of job classes falling in each are shown in Table 3. Job family 20 includes those jobs which are highest on any one of three factors: Factor I, filing, sorting, routing, and related activities: Factor IV, typing or key data entry; and Factor VII, composing and editing. (Stenography did not appear as a primary factor, but all jobs involving significant stenographic tasks grouped under Factor I and are included in Job Family 20.)

Table 2

Identified Factors and Number of Job Classes Loading Highest on Each

	Factor	No. of Classes	%
I.	Filing, sorting, routing, and related activities	80	46
II.	Bookkeeping and financial activities	41	23
III.	Interacting with people	35	20
IV.	Typing, or key data entry	11	6
V.	Handling materials	4	2
VI.	Computing and coding activities	2	1
VII.	Composing and editing activities	1	1
VIII.	Supervising, directing, and deciding	0	0
IX.	Operating Machines	1	1
	Total	175	100

Note: All job classes loaded .50 or greater on the factor for which they are listed.

Table 3

Number of Jobs Belonging to Each D.O.T. Clerical Job Family
(Based on Highest Factor Loading of Each Job)

	Factor									
	I	II	III	IV	V	VI	VII	IX		
D.O.T. Job Family Code	Filing/ Sorting	Bookkeeping/ Financial	Interacting w/people	Typing/ Key Data	Handling Materials	Computing/ Coding	Composing/ Editing	Operating-Machines	Total n	Jobs %
20	80			11			1		92	53
21		41				2			43	25
22					4				4	2
23			35						35	20
24								1	1	1
								Total	175	100

Notes: (1) All jobs had a factor loading of at least .50 on the factor under which they are included.
(2) D.O.T. Codes (4th Edition):
20 - Steno, Typing, Filing and Related Occupations
21 - Computing and Account Recording
22 - Production and Stock Clerks
23 - Information and Message Distribution
24 - Miscellaneous Clerical Occupations

Job family 21 includes those jobs which are highest on Factor II, bookkeeping and financial activities and Factor VI, computing and coding activities. Job family 22 covers those jobs which are highest on Factor V, handling materials. Job family 23 includes those jobs highest on Factor III, interacting with people. Only one job, Microfilm Operator-Records Clerk, was included in job family 24, Miscellaneous.

A mean profile across job classes was computed for each primary job type derived from factor analysis. These profiles linked to D.O.T. job families are shown in Table 4. As might be expected the differences in job types are primarily differences of degree. Each job type emphasizes one or two of the functions (duties) relative to the others.

Table 4

Mean Duty Profiles (% Time Spent) of Primary Job Types Under D.O.T. Families

	D. O. T. Clerical Family							
		20		21		22	23	24
Duty	Factor I Duty E	Factor VI Duty M	Factor VII Duty C	Factor II Duty B	Factor VI Duty D	Factor V Duty G	Factor III Duty H	Factor IX Duty I
A. Arranging for appointments, events	3	1	6	1	0	2	2	0
B. Bookkeeping and Financial	5	7	5	28	8	9	5	0
C. Composing or Editing	8	9	27	3	1	7	5	7
D. Computing and Coding	4	6	2	8	26	4	3	5
E. Filing, Sorting, Routing	21	10	9	11	19	3	12	28
F. Gathering Information	2	2	0	1	1	1	2	0
G. Handling Material	9	8	12	6	4	26	7	23
H. Interacting with people	16	8	20	14	9	15	35	4
I. Operating Machines	5	6	7	6	4	6	5	27
J. Maintaining, Processing records	11	9	0	11	10	18	10	0
K. Supervising, Directing, Deciding	4	3	0	4	4	4	5	0
L. Stenography	2	0	0	0	0	0	1	0
M. Typing/Key Data Entry	9	31	5	5	14	5	5	5
N. Miscellaneous	2	1	6	2	0	1	2	0
	101%	101%	99%	100%	100%	101%	99%	99%

DISCUSSION

For the purpose of this study the CODAP cluster analysis failed to yield clear-cut results. The Q-type factor analysis based on composite CODAP job descriptions provided job types which could be linked clearly to D.O.T. job family categories. The D.O.T. families, in turn, are linked to a substantial data base of validities for specific test types, including measures of the type used in the Federal clerical tests. Special skills requirements such as typing and stenography, although not identified from the factor analysis, are clearly specifiable from the CODAP job descriptions for job classes or individuals.

Job classes cannot be recommended for inclusion in cooperative examining simply on the basis of their job family grouping. Each job class must be reviewed separately with respect to other variables, e.g., extent to which incumbents believe the task inventory to adequately cover their job, adequacy of sampling for the job class, and appropriateness of filling the job at or near entry level.

The study should be replicated in another geographical area. Given similar results, particularly for the Federal job classes, the profiles can be cumulated. Job analysis for potential cooperative examining can then be simplified by administering the task inventory and matching the profile for the job class to the most similar "standard" profile, assuming other relevant considerations are supportive.

REFERENCE NOTES

1. Pearlman, K. and Schmidt, F. L. Test of a new model of validity generalization: results for tests used in clerical selection. Paper presented at annual conference of the International Personnel Management Association Assessment Council, Atlanta, Ga., June 26-29, 1978.

2. Christal, R. E. The United States Air Force occupational research project. (AFHRL-TR-73-75) Brooks AFB, Texas: Air Force System Command, 1974.

REFERENCES

Blashfield, R. K. Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchical methods. Psychological Bulletin, 1976, Vol. 83, No. 3, 377-388.

Overall, J. E. and Klett, C. J. Applied multivariate analysis. New York: McGraw-Hill, 1972.

Schmidt, F. L. and Hunter, J. E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

U. S. Department of Labor. Dictionary of occupational titles. (4th ed.) Washington. D.C.: U.S. Government Printing Office, 1977.

U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U. S. Department of Justice, and U. S. Department of Labor. Uniform guidelines on employee selection procedures (1978). Washington, D. C.: 43 Federal Register No. 166, August 25, 1978.

Ward, J. H., Jr. Hierarchical grouping to optimize an objective function. American Statistical Association Journal, 1963, vol. 58, 236-244.

MILITARY PAY: IS IT COMPARABLE?*

Linda D. Pappas

General Research Corporation
McLean, Virginia 22102

*Henry Alberts, Karen Alderman, John Finneran, Alvin Frost, Janice Hillman, Roger Johnson, Bertil Lofstrom, Robin Lovely, and Thomas Wilson participated in the co-authorship of this paper.

MILITARY PAY: IS IT COMPARABLE?[1]

Linda D. Pappas

General Research Corporation
McLean, Virginia 22102

INTRODUCTION AND BACKGROUND

The adequacy of United States military compensation is a matter of continuing concern. Elimination of the draft, coupled with the impending manpower shortages in the mid-1980's, Federal pay caps and continuing budget constraints, require the most judicious employment of compensation as a major management tool.

Numerous compensation studies have been conducted during the past 30 years.[2] However, until the early 1960's, military compensation levels lagged significantly behind those of the private sector.

- Basic pay was not increased from 1922 to 1940.
- From 1946 to 1963, five military raises increased pay by 76%. During that same time frame, nine Classification Act raises increased civil service pay by 125% and nine Postal Field Service raises increased postal pay by 177%.
- Special and incentive pays and allowances did not keep pace with inflation nor did they retain their original percentage relationship to base pay.
- Until 1972, little attention was paid to the financial needs of new recruits because of existence of the draft.

As noted in a 1974 Rand Corporation report:

> There was no apparent need to offer inducements to attract more or better qualified young men to military service; first term

[1] Data from: Selected Military Compensation Issues, Linda Pappas, et. al. General Research Corporation, prepared for Chief of Naval Personnel, September 1979.

[2] Hook Commission (1948); Cordiner Committee (1957); Gorham Committee and Randall Panel (1962); Folsom Panel (1965); First, Second, and Third Quadrennial Reviews of Military Compensation (QRMC) (1967, 1971, 1977); and the President's Commission on Military Compensation (1978). In addition, compensation was a major issue for the Gates Commission (1969) and the Defense Manpower Commission (1976).

> accessions were treated with what was characterized as "deliberate neglect...." For example, from 1952 to 1963 enlisted men with fewer than two years of service received no pay increases whatever....[1]

The situation changed somewhat in the next decade (1963-1972):

> Within the past 10 yeais, the following radical changes have, in fact, occurred in the military compensation system: comparability pay increases (1965-1970); special pay programs, such as the variable reenlistment bonus (1966) and the combat enlistment bonus (1972); and the AVF pay increases for first-termers (1969-1971).[2]

Despite the attainment of general military/civilian comparability with respect to basic pay and allowances in 1972, there has been a steady erosion of benefits since then. This erosion has been manifested in many areas: loss of money (termination of superior performance pay, reallocation of pay increases, elimination of regular reenlistment bonuses, etc.); decreased medical benefits (CHAMPUS changes, reduction of hospital services, etc.); decreased education benefits (elimination of the fully funded GI Bill, etc.); and also decreases in retirement benefits (elimination of 1% kicker, PCMC recommendation of annuity at age 60 or 30 years of service).

Current attraction and retention problems in the military suggest an examination of compensation is necessary. Specifically, one compensation approach of importance is pay comparability or pay adequacy.

APPROACHES TO COMPARABILITY

Four major approaches in pay comparability are examined in this paper: age earning profiles, work level comparisons, compensation paid for specific jobs, and lifestream earnings. Age earning profiles were derived for the general population and compared with military age earnings. Comparisons were made between military and the Civil Service General Schedule (white collar) and Wage Grade (blue collar) compensation at select grade linkages. This provided work-level difficulty comparisons. Compensation paid to select jobs (e.g., aviators, firemen, police) in the private sector was examined and compared to military pay for the same or similar jobs. Additionally, lifestream earnings comparisons were developed for certain of these select jobs.

[1] F. J. Morgan and D. E. Roseen, Recruiting, Classification and Assignment in the All-Volunteer Force: Underlying Influence and Emerging Issues, Report No. R-1357-ARPA, The Rand Corporation, Santa Monica, California, June 1974.

[2] D. L. Jaquette and G. R. Nelson, The Implications of Manpower Supply and Productivity for the Pay and Composition of the Military Force: An Optimization Model, Report No. R-1451-ARPA, The Rand Corporation, Santa Monica, California, July 1974.

It is noteworthy that, in employing the preceding approaches, the more general approaches (e.g., age earning profiles) tend to show the military relatively well paid compared to the nonmilitary sector; the more specific the approaches (e.g., lifestream earnings for select occupations) shows that the military pay somewhat lags that of the nonmilitary sector.

METHOD

Compensation Components Defined

Military Compensation

Compensation was examined with respect to RMC and its components, base pay, basic allowance for subsistence (BAS), basic allowance for quarters (BAQ), and tax advantage (TAD).

Other differentials, i.e., special pays recognizing compensation for particular work locations or environments, family separation, job hazards, or skills were not considered.

Private Sector Compensation

Efforts were made, insofar as data permitted, to differentiate between basic pay and special differentials in the private sector. Some large organizations use an equivalent differential compensation similar to BAQ to remunerate employees for extra costs of housing when they are on an extended away-from-home assignment. Another technique is increasing an employee's base pay to compensate for differences in the cost of living at different geographic locations within an employer's organization. Some organizations reimburse additional cost directly as an expense, using the expense voucher technique. Still others use meals as payment in kind for normal work situations if no alternative to use of the company-operated cafeteria is practical. An example of private sector use of "payment in kind" is the extensive lodging and messing facilities established by the Alaska Pipeline Consortium of oil companies specifically for their employees. The benefits of food and lodging were made in kind because no alternative existed. Base pays for workers receiving compensation in kind for food and housing were not reduced over those received by employees in urban areas where food and housing were available.[1] It appears that housing and food were considered offsets for "place in which work was performed."

Time Frame Defined

The central focus of this study is for the years 1972-1978. Temporal relationships were examined over this period. The 1972 starting point was selected because pay comparability between the military and

Federal Civil Service was defined to exist at that time in an Office of the Secretary of Defense (OSD) report submitted to Congress.

COMPARABILITY APPROACHES

Age Earning Profiles

The first of the four quantitative approaches employed was that of age earning profiles (AEP). In this approach, populations are grouped by selecting cohorts of like age groups. These groupings are then linked to salaries or wages earned. A plot of the data generates curves that are commonly called age earning profiles or age-education earning profiles. Using the total census population and the census white male population, plots for 1976 are presented in Figure 1.

A major strength of this method is the availability of data. However, the weakness is the lack of selectivity of the nonmilitary cohort. Such factors as productivity, type of work, working conditions, personnel systems, and quality of work are not considered.

Work-Level Comparisons

The quantitative work-level comparison approach is more specific than AEPs but more general than specific job comparisons. This approach employs job analysis and equates jobs of similar work-level difficulty by grade (e.g., GS-5 and E-5). Work-level comparisons are defined as they were in the 3rd QRMC and the PCMC. Linkages are based on substantially equal job content and difficulty between select military and Civil Service General Schedule (GS) and Wage Grade (WG) grades. This approach employs the concept of substantially equal pay for substantially equal work. It should be noted that the new Civil Service Reform Act, October 13, 1978, PL 95-454, states the pay comparability principle as:

> Equal pay should be provided for work of equal value, with appropriate consideration of both national and local rates paid by employers in the private sector, and appropriate incentives and recognition should be provided for excellence in performance.

Pay Bands

Pay bands were developed to compare military pay (RMC) to Civil Service blue collar (WG) pay as well as white collar pay for both the private sector[1] and Civil Service. The pay bands express the population distribution by income at selected grades. The pay band comparison was accomplished for both enlisted personnel (Figure 2) and officer personnel (Figure 3).

[1] Data for the private sector were derived from the Professional, Administrative, Technical and Clerical (PATC) Survey performed by the Bureau of Labor Statistics to identify the appropriate amount of the annual Civil Service comparability pay raise.

FIGURE 1

1976 CENSUS AND MILITARY AGE EARNINGS

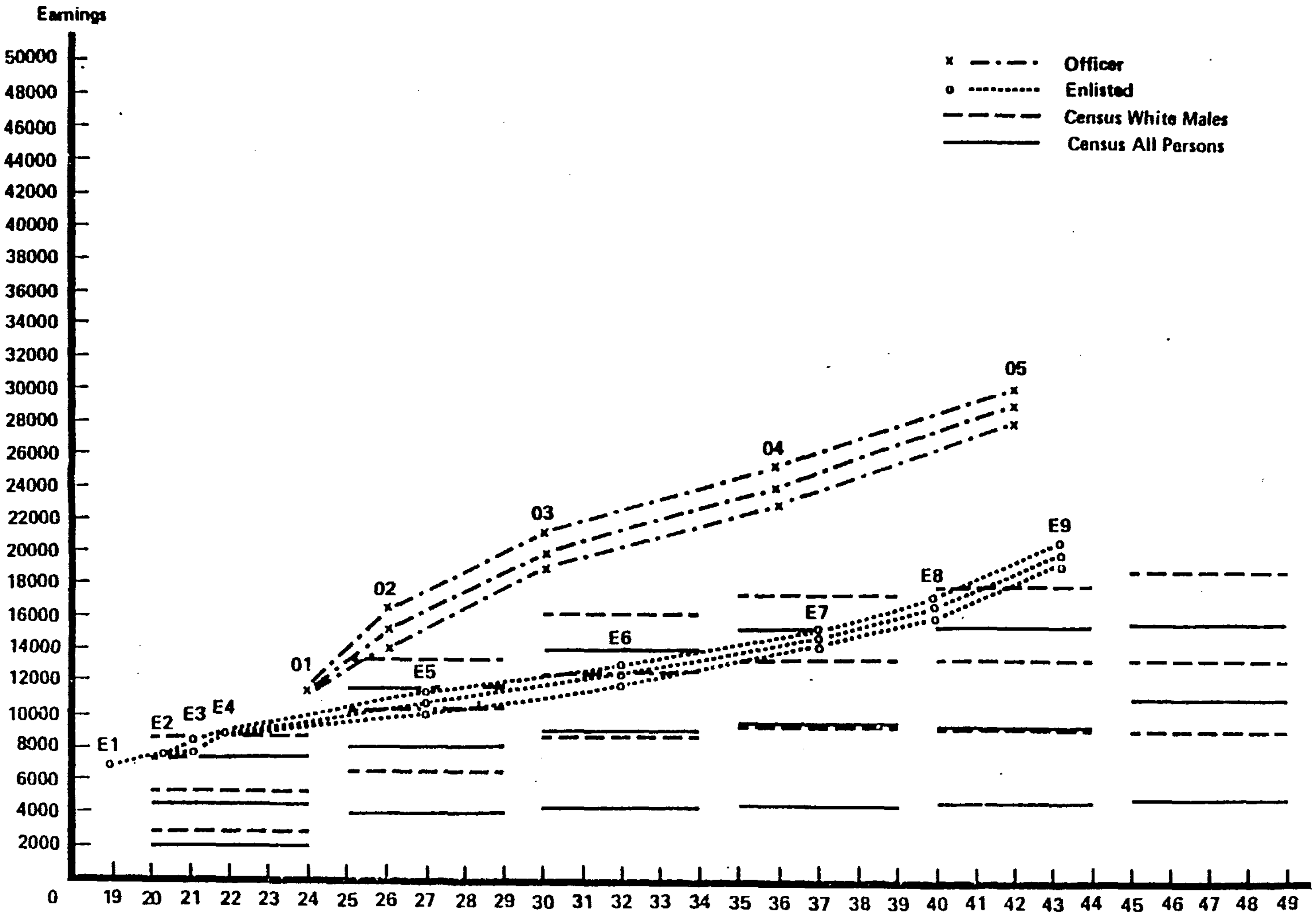
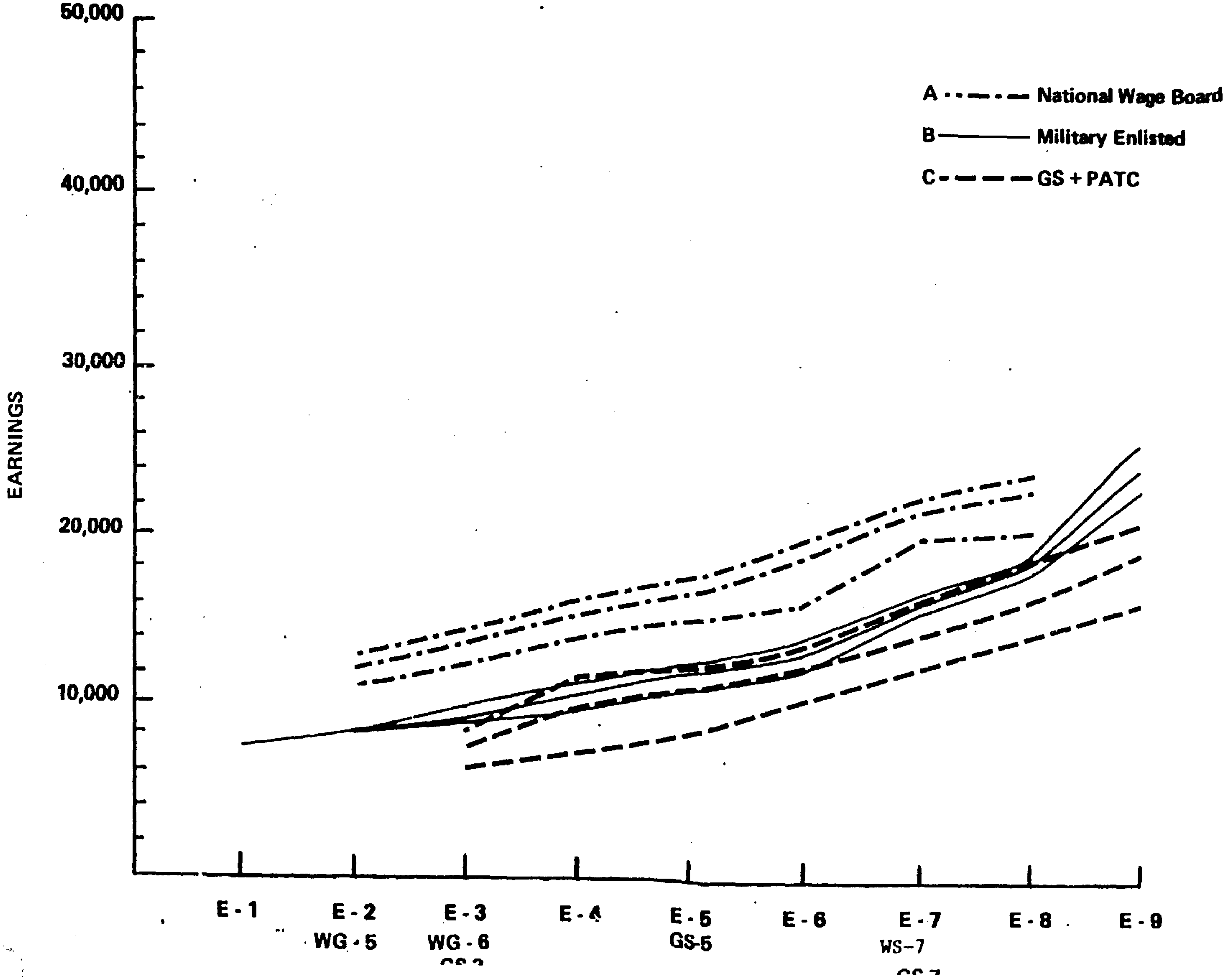
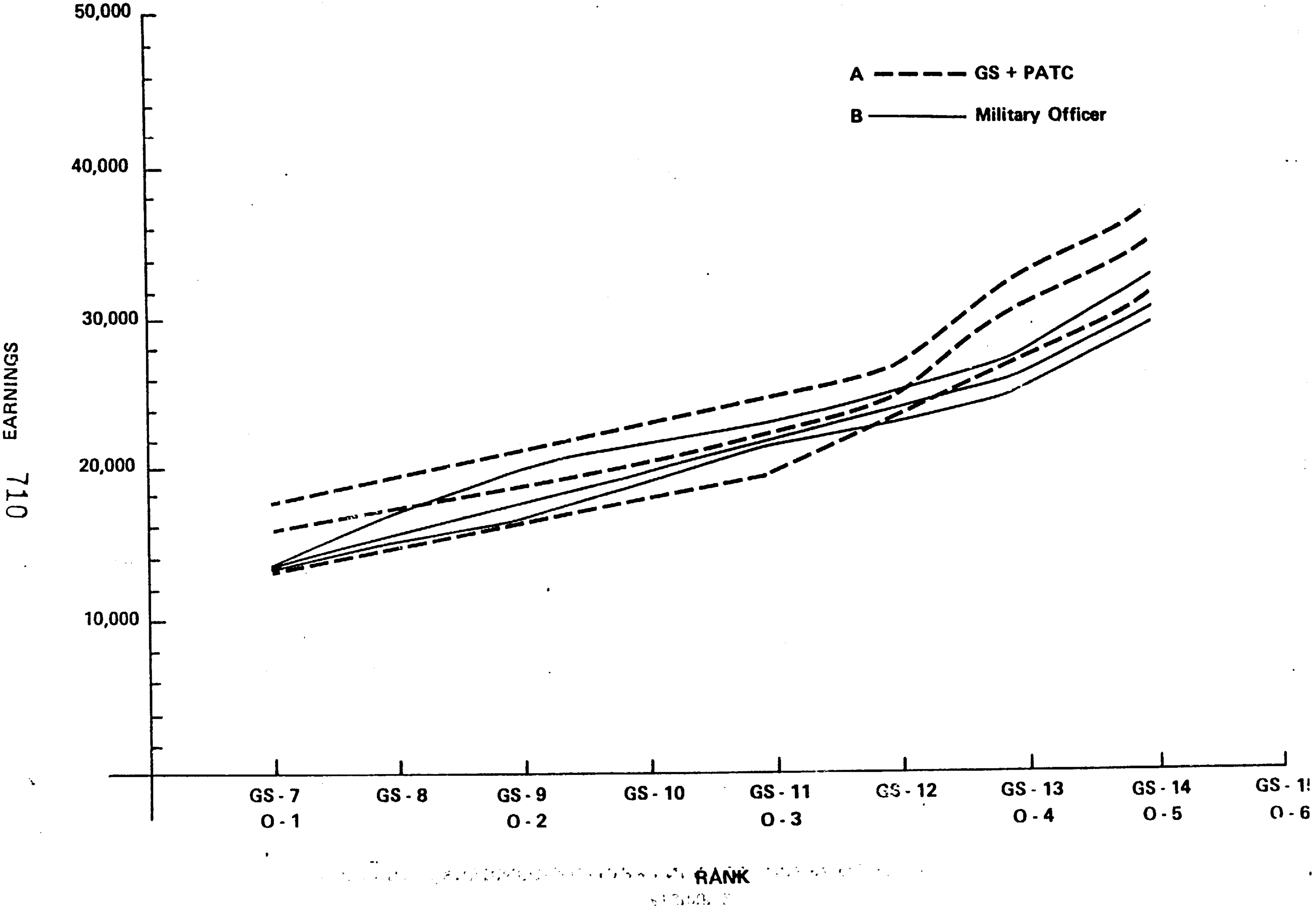


FIGURE 2

1977 GS + PATC COMBINED; MILITARY ENLISTED; AND NATIONAL WAGE GRADE PAY BANDS
(GS - EQUIVALENT GRADES 1-7)



(GS- EQUIVALENT GRADES 7-15)



In general, all quartiles of the enlisted RMC pay band lagged those of the blue collar WG cohort. With respect to white collar workers, only the first quartile of the enlisted pay band for grades E-4 through E-6 lagged that cohort's second quartile (median). This is important since the pay lag coincides with enlistment decision points to remain in or leave the service. Officer pay for all grades falls well below the third quartile of the GS/PATC comparison cohort (Figure 3) and for O-1 through O-6 the first two quartiles of the military pay band fall below the second quartile (and in some cases the first quartile) of the GS/PATC cohort.

Selected Occupations and Selected Lifestream Earnings

Selected occupations and selected lifestream earnings are the two most specific comparability approaches examined. Compensation for specific jobs is examined as well as lifestream earnings (LSE). Specific military earnings streams are compared with specific civilian job earnings streams in cases where such comparisons appeared useful. The strength of these approaches is the specificity of comparison; a limitation is that these approaches cannot be applied to all military jobs.

Enlisted Personnel

Police and Firefighters. Earnings for police and firefighters were developed and compared with military earnings. Because there are no special pays for general military service and no special pays apply to police and firefighters, it was felt that some direct relationship might be adduced. Military persons keep the peace in a special sense, and they perform internationally. Police and firefighters keep the peace within more restricted political boundaries.

Table 1 presents earnings in 1977 for police and firefighters in four selected locations. It can be seen that the civilian peace-keeping and fire-protection organizations compensate their rank and file at higher levels than do the military forces. Figure 4 graphically compares Washington, D.C. police/firefighters pay to enlisted military pay.

Lifestream earnings for the police and firefighters are illustrated in Tables 2 and 3. Certain assumptions are made about the velocity through the various grade structures.

Foreman/Engineer at a Power Station. An internal Naval document comparing earnings for a Navy BT2 and a fireman/engineer at the San Diego Gas and Electric Company was reviewed. There is a recognition pay included in BT2 compensation, but the position of foreman/engineer has no special recognition pay associated with its performance.

The information presented evaluates the particular 8 years-of-service point for both individuals. There is insufficient information available to make a lifestream earnings comparison. However, a single point on the lifestream earnings continuum can be compared. This comparison reveals

TABLE 1

COMPENSATION FOR POLICE AND FIREFIGHTERS
(1977)

Rank	Mass State Police	Calif Hwy Patrol	Boston Municipal Firefighter/Police	Washington, D.C. Firefighter/Police
Recruit	6,600	15,108-18,036	-	
Patrolman/Trooper/ Firefighter	10,620-17,820	16,512-18,864	13,105-17,064	13,799-19,871
Next Supervisor	12,300-19,680	18,444-22,116	20,522-	15,732-21,525
Next Supervisor	13,560-21,720	21,132-25,356	23,601-	17,248-22,423
Next Supervisor	14,940-23,940	25,356-30,504	27,140-	18,741-23,428
Next Supervisor	16,500-26,340	29,124-35,064	-	21,662-25,997
Next Supervisor		31,932-38,508	-	23,667-27,205
Next Supervisor		30,708-39,720		25,665-29,513
Next Supervisor				29,750-34,223
Next Supervisor				34,913-42,009
Asst Chief of Police/ Asst Fire Chief			31,711	41,396-46,915
Deputy Commissioner	27,000	38,040		
Police Chief/Fire Chief				47,879-50,781
Commissioner	31,000	40,764		

FIGURE 4

1977 GS + PATC COMBINED AND MILITARY ENLISTED PAY BANDS
(GS EQUIVALENT GRADES 1-7)

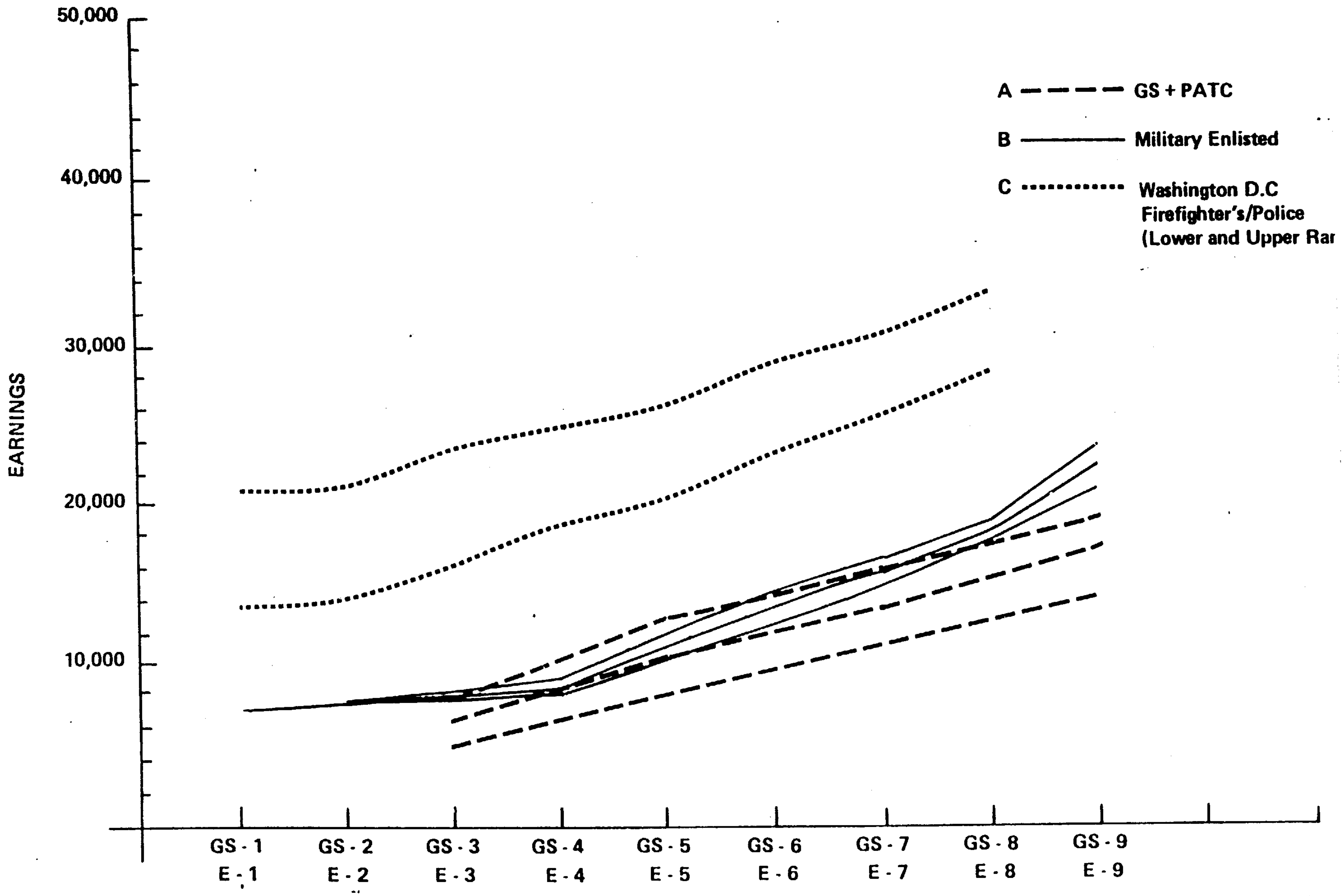


TABLE 2

LIFESTREAM EARNINGS OF AN 0-5 RETIRING ON 10-1-76
PAY RATE, AGE 41, 20 YEARS OF SERVICE AND A D.C. POLICE LIEUTENANT
RETIRING IN 1976, AGE 40, 20 YEARS OF SERVICE

[Military earnings, retirement pay, post-retirement income, Social Security (PIA)*]

	0-5	Police Lieutenant
Career pay	$ 220,364	$ 224,251
Retired pay to age 65	300,936	292,692
Subtotal	521,300	516,943
Post retirement income: $20K, 24 yrs.	480,000	480,000
Subtotal	1,001,300	996,943
Retired pay age 66-75	125,390	121,955
Subtotal	1,126,690	1,118,898
Soc. Sec. (PIA) Individual	30,120	35,472
GRAND TOTAL	1,156,810	1,154,370

TABLE 3

LIFESTREAM EARNINGS OF AN E-7, RETIRING AGE 39, 20 YEARS OF SERVICE, 1976
LIFESTREAM EARNINGS OF A WASHINGTON, D.C., FIREFIGHTER,
RETIRING IN 1976, AGE 41, 20 YEARS OF SERVICE
[Earnings, retirement income, post-retirement income, Social Security (PIA)*]

	E-7	Firefighter
Career pay	143,622	200,353
Retired pay to age 65	141,991	230,496
Subtotal	285,613	430,849
Post retirement income: $9K, 24 yrs.	225,000 (25 yrs.)	216,000 (24 yrs.)
Subtotal	510,613	646,849
Retired pay age 66-75	54,612	96,040
Subtotal	565,225	742,889
Soc. Sec. (PIA)		
Individual	27,156	35,472
Family (max.)		
GRAND TOTAL	592,381	778,361

*PIA = Primary Insurance Amount

that, at the 8-year point, one year's earnings of the fireman/engineer are $22,690. The BT2 earns a compensation of approximately $15,130 including his selective reenlistment bonus.

The fireman/engineer receives more compensation for performance of his ordinary work task than does the BT2. The fireman/engineer has ordinary amounts of overtime included in his $22,690 compensation. The BT2 has a minimum 40-hour workweek and may work up to 72 hours a week when his ship is at sea.

SUMMARY OF FINDINGS

The comparisons made between military and civilian compensation levels for age earning cohorts, linked grades, and job-by-job career paths indicate that:

- When stratified only by age, enlisted military personnel do not receive consistently higher or lower incomes than their nonmilitary cohorts. Military officers tend to receive higher incomes than their nonmilitary cohort.

- When work-level comparisons are employed, pay for military personnel, both officer and enlisted sometimes lags and sometimes exceeds that of its comparison cohort.

- When compared on the basis of similar skill levels and career paths, military personnel tend to receive lower compensation than their civilian cohorts. Differences may be significant. This is especially noteworthy because of the special skill levels demanded of many military incumbents.

In conclusion, when attempting to ascertain military pay comparability, the cohort selected for comparison and the methodology must be chosen carefully in order to reflect a fair picture.

RESEARCH AND MANAGEMENT APPLICATIONS OF THE COMPUTER ASSISTED REFERENCE LOCATOR (CARL) SYSTEM[1]

William A. Sands
Loralee Hartmann

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

Background

The Computer Assisted Reference Locator (CARL) System is a computer-based interactive information retrieval system which was introduced at the 20th Annual Military Testing Association (MTA) Conference (Sands, 1978). The approach employed in the system, termed "coordinate indexing," was mentioned in the psychological literature by Broadhurst (1962). This approach involves assigning a sequential number to each document as it is incorporated into the system and identifying those keyterms which characterize the document. Subsequent interactive retrieval of target documents is based upon these keyterms, each of which can be considered as one of a set of classification coordinates.

The previously mentioned paper on the CARL System provided a brief review of selected information retrieval systems.[2] In addition, this overview covered the design objectives, the way in which information is encoded for input to the system, and the individual data files and computer programs which make up the CARL System.

Purposes

The present paper will briefly review the CARL System and then focus on the QUERY program. This interactive computer program is the part of the CARL System with which most users have contact. The examples employed will illustrate some of the potential applications of the CARL System in both research and management.

CARL SYSTEM

Design Objectives

Three objectives were considered in designing the CARL System: (1) simplicity of reference query and retrieval, (2) ease of system maintenance, and (3) adaptability for alternative computer systems.

Data Files

The CARL System contains five data files: (1) a user dictionary file (UDXXXX.),[3] (2) an original data file (ODXXXX.), (3) an inverted keyterm file (IKXXXX.), (4) a keyword dictionary file (KDXXXX.), and (5) an author dictionary file (ADXXXX.).

The user dictionary file (Figure 1) contains, in alphabetical order, all words which can be employed in a reference search.[4] In addition to providing a list of all legitimate keywords, the user dictionary provides guidance on the form of the keyword that should be used. If, for example, the user wishes to retrieve all references dealing with "assessment," then examination of the user dictionary would indicate that "ASSESS" is the appropriate keyterm. The user dictionary also defines all permissible abbreviations (e.g.. "NPRDC" is the keyterm which will retrieve all reports in the system which have been published by the Navy Personnel Research and Development Center). Finally, the user dictionary provides guidance on synonyms (e.g., "METHOD" should be used instead of "procedure").

The original data file contains the raw data for all references in ascending numerical sequence. Four types of formats are used in the original data file: (a) header type (card #Ø); (b) reference citation type (cards #1-6, as necessary); (c) author(s) type (card(s) #7, as necessary); and (d) keyword(s) type (card(s) #8, as necessary). Figure 2 presents an example of an original data file containing four references.

The inverted keyterm file contains an alphabetical listing of keyterms (i.e., keywords and authors) and the reference numbers associated with them. Figure 3 illustrates the inverted keyterm file corresponding to the original data file shown in Figure 2.

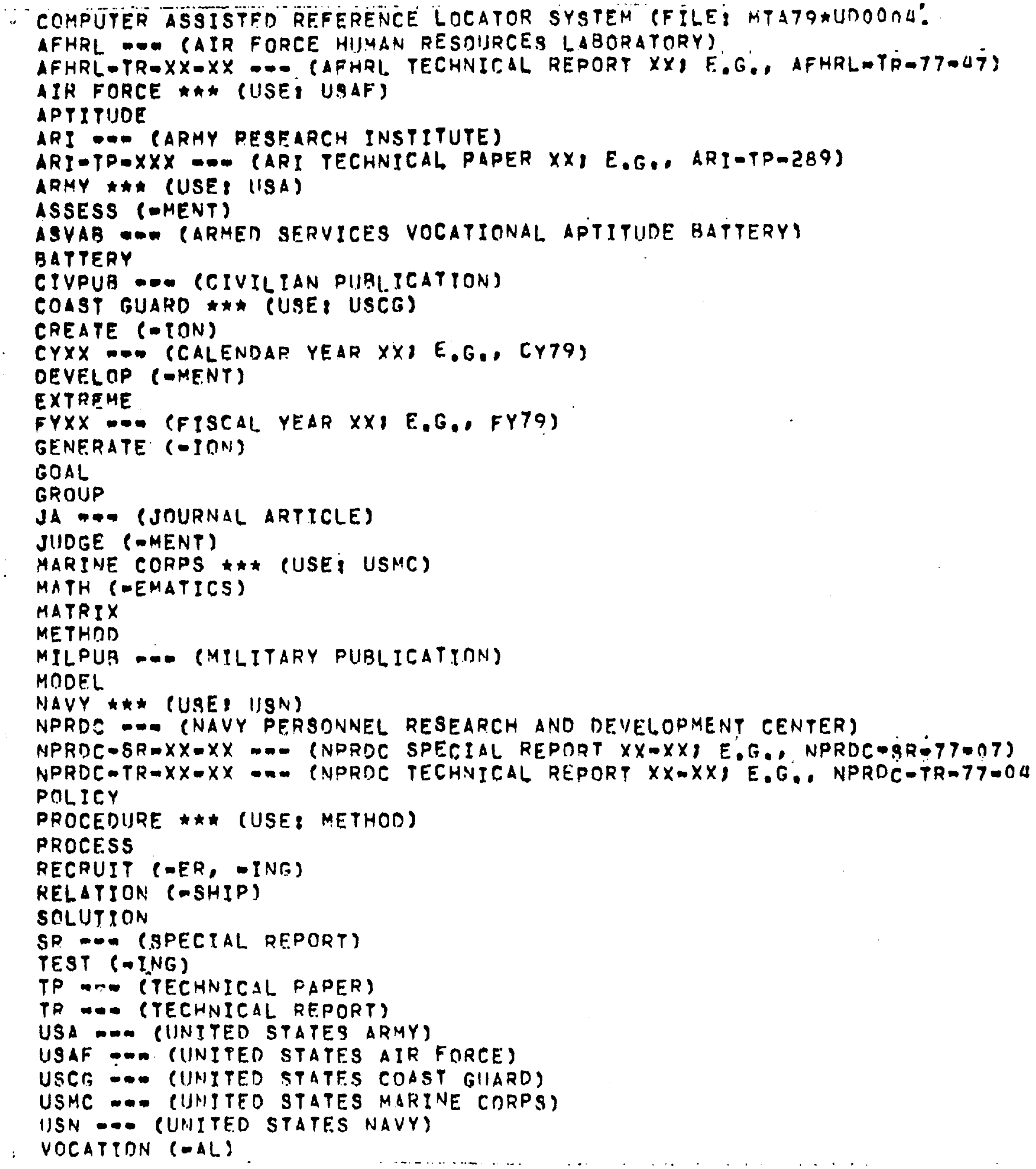
The keyword dictionary file contains, in alphabetical order, all the keywords which have been employed to characterize the references in the system. Similarly, the author dictionary contains, in alphabetical sequence, all the authors of the references in the system.

Computer Programs

The CARL System contains various computer programs which may be divided into two categories: (1) input preparation programs and (2) system programs. There are three input preparation programs: (1) an editing program (EDIT), (2) a recommended keyword program (RKWD), and (3) a duplicate identification program (DUPL). There are four system programs: (1) a system file generation

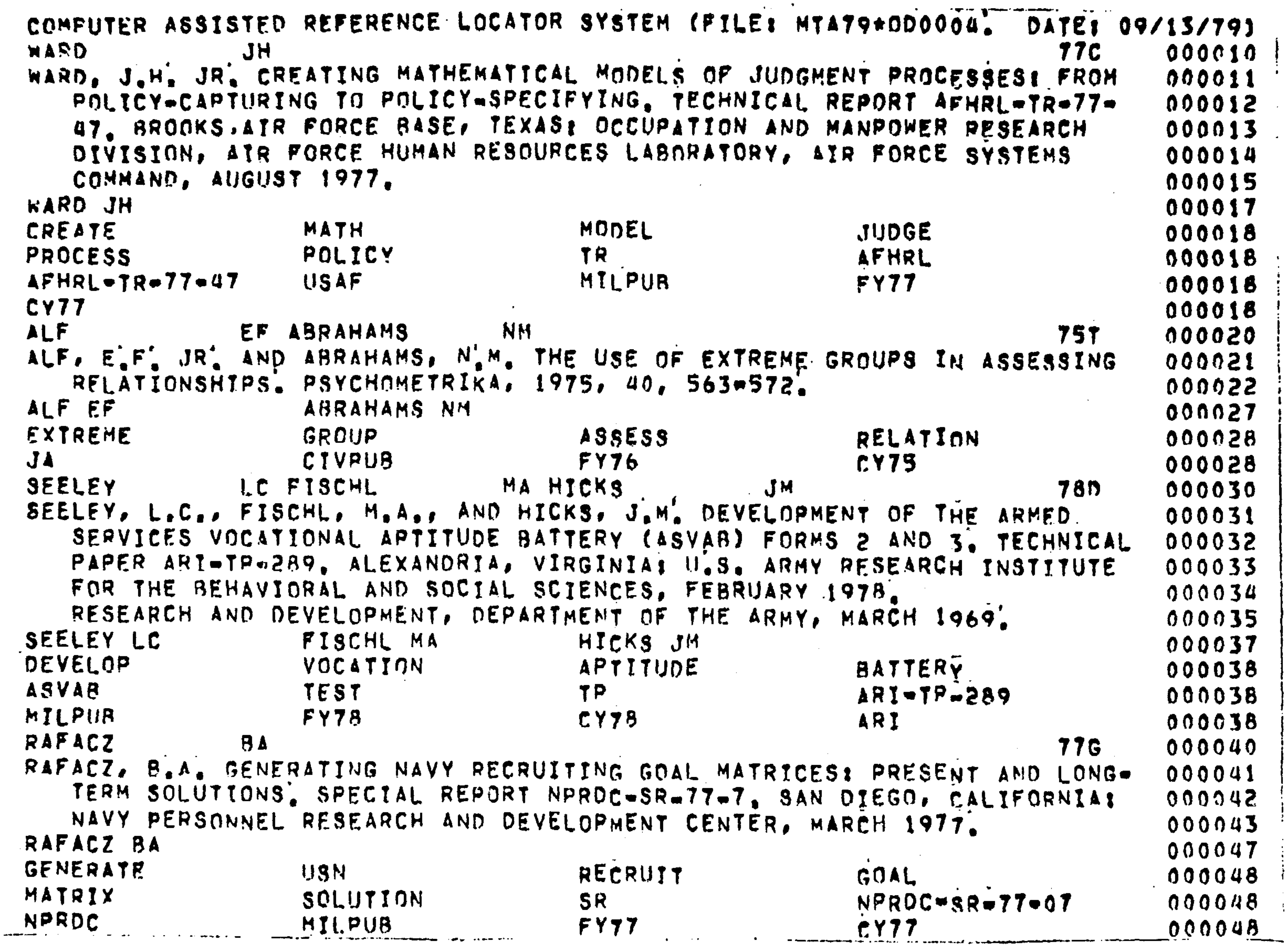
[3] The XXXX portion of each data file name is a number indicating the highest reference number currently incorporated into the system. For example, UD2ØØØ. would be the name of the user dictionary file for references 1-2ØØØ, inclusive.

[4] While not included in the user dictionary, author names may also be used as keyterms for queries.



```
COMPUTER ASSISTED REFERENCE LOCATOR SYSTEM (FILE: MTA79*UD0004.
AFHRL --- (AIR FORCE HUMAN RESOURCES LABORATORY)
AFHRL-TR-XX-XX --- (AFHRL TECHNICAL REPORT XX; E.G., AFHRL-TR-77-47)
AIR FORCE *** (USE: USAF)
APTITUDE
ARI --- (ARMY RESEARCH INSTITUTE)
ARI-TP-XXX --- (ARI TECHNICAL PAPER XX; E.G., ARI-TP-289)
ARMY *** (USE: USA)
ASSESS (-MENT)
ASVAB --- (ARMED SERVICES VOCATIONAL APTITUDE BATTERY)
BATTERY
CIVPUB --- (CIVILIAN PUBLICATION)
COAST GUARD *** (USE: USCG)
CREATE (-ION)
CYXX --- (CALENDAR YEAR XX; E.G., CY79)
DEVELOP (-MENT)
EXTREME
FYXX --- (FISCAL YEAR XX; E.G., FY79)
GENERATE (-ION)
GOAL
GROUP
JA --- (JOURNAL ARTICLE)
JUDGE (-MENT)
MARINE CORPS *** (USE: USMC)
MATH (-EMATICS)
MATRIX
METHOD
MILPUB --- (MILITARY PUBLICATION)
MODEL
NAVY *** (USE: USN)
NPRDC --- (NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER)
NPRDC-SR-XX-XX --- (NPRDC SPECIAL REPORT XX-XX; E.G., NPRDC-SR-77-07)
NPRDC-TR-XX-XX --- (NPRDC TECHNICAL REPORT XX-XX; E.G., NPRDC-TR-77-04
POLICY
PROCEDURE *** (USE: METHOD)
PROCESS
RECRUIT (-ER, -ING)
RELATION (-SHIP)
SOLUTION
SR --- (SPECIAL REPORT)
TEST (-ING)
TP --- (TECHNICAL PAPER)
TR --- (TECHNICAL REPORT)
USA --- (UNITED STATES ARMY)
USAF --- (UNITED STATES AIR FORCE)
USCG --- (UNITED STATES COAST GUARD)
USMC --- (UNITED STATES MARINE CORPS)
USN --- (UNITED STATES NAVY)
VOCATION (-AL)
```

Figure 1. User Dictionary File (UDØØØ4.).



```
COMPUTER ASSISTED REFERENCE LOCATOR SYSTEM (FILE: MTA79*OD0004.  DATE: 09/13/79)
WARD          JH                                                        77C    000010
WARD, J.H. JR. CREATING MATHEMATICAL MODELS OF JUDGMENT PROCESSES: FROM        000011
   POLICY-CAPTURING TO POLICY-SPECIFYING, TECHNICAL REPORT AFHRL-TR-77-        000012
   47, BROOKS AIR FORCE BASE, TEXAS: OCCUPATION AND MANPOWER RESEARCH          000013
   DIVISION, AIR FORCE HUMAN RESOURCES LABORATORY, AIR FORCE SYSTEMS           000014
   COMMAND, AUGUST 1977.                                                       000015
WARD JH                                                                        000017
CREATE              MATH                MODEL               JUDGE              000018
PROCESS             POLICY              TR                  AFHRL              000018
AFHRL-TR-77-47      USAF                MILPUR              FY77               000018
CY77                                                                           000018
ALF           EF ABRAHAMS       NM                                      75T    000020
ALF, E.F. JR. AND ABRAHAMS, N.M. THE USE OF EXTREME GROUPS IN ASSESSING        000021
   RELATIONSHIPS. PSYCHOMETRIKA, 1975, 40, 563-572.                            000022
ALF EF              ABRAHAMS NM                                                000027
EXTREME             GROUP               ASSESS              RELATION           000028
JA                  CIVPUB              FY76                CY75               000028
SEELEY        LC FISCHL        MA HICKS           JM                    78D    000030
SEELEY, L.C., FISCHL, M.A., AND HICKS, J.M. DEVELOPMENT OF THE ARMED           000031
   SERVICES VOCATIONAL APTITUDE BATTERY (ASVAB) FORMS 2 AND 3. TECHNICAL       000032
   PAPER ARI-TP-289, ALEXANDRIA, VIRGINIA: U.S. ARMY RESEARCH INSTITUTE        000033
   FOR THE BEHAVIORAL AND SOCIAL SCIENCES, FEBRUARY 1978.                      000034
   RESEARCH AND DEVELOPMENT, DEPARTMENT OF THE ARMY, MARCH 1969.               000035
SEELEY LC           FISCHL MA           HICKS JM                               000037
DEVELOP             VOCATION            APTITUDE            BATTERY            000038
ASVAB               TEST                TP                  ARI-TP-289         000038
MILPUR              FY78                CY78                ARI                000038
RAFACZ        BA                                                        77G    000040
RAFACZ, B.A. GENERATING NAVY RECRUITING GOAL MATRICES: PRESENT AND LONG-       000041
   TERM SOLUTIONS. SPECIAL REPORT NPRDC-SR-77-7, SAN DIEGO, CALIFORNIA:        000042
   NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER, MARCH 1977.                 000043
RAFACZ BA                                                                      000047
GENERATE            USN                 RECRUIT             GOAL               000048
MATRIX              SOLUTION            SR                  NPRDC-SR-77-07     000048
NPRDC               MILPUB              FY77                CY77               000048
```

Figure 2. Original Data File (ODØØØ4.).

```
COMPUTER ASSISTED REFERENCE LOCATOR SYSTEM (FILE: MTA79*IK0004.
ABRAHAMS NM            2    0    0    0    0
AFHRL                  1    0    0    0    0
AFHRL-TR-77-47         1    0    0    0    0
ALF EF                 2    0    0    0    0
APTITUDE               3    0    0    0    0
ARI                    3    0    0    0    0
ARI-TP-289             3    0    0    0    0
ASSESS                 2    0    0    0    0
ASVAB                  3    0    0    0    0
BATTERY                3    0    0    0    0
CIVPUB                 2    0    0    0    0
CREATE                 1    0    0    0    0
CY75                   2    0    0    0    0
CY77                   1    4    0    0    0
CY78                   3    0    0    0    0
DEVELOP                3    0    0    0    0
EXTREME                2    0    0    0    0
FISCHL MA              3    0    0    0    0
FY76                   2    0    0    0    0
FY77                   1    4    0    0    0
FY78                   3    0    0    0    0
GENERATE               4    0    0    0    0
GOAL                   4    0    0    0    0
GROUP                  2    0    0    0    0
HICKS JM               3    0    0    0    0
JA                     2    0    0    0    0
JUDGE                  1    0    0    0    0
MATH                   1    0    0    0    0
MATRIX                 4    0    0    0    0
MILPUB                 1    3    4    0    0
MODEL                  1    0    0    0    0
NPRDC                  4    0    0    0    0
NPRDC-SR-77-07         4    0    0    0    0
POLICY                 1    0    0    0    0
PROCESS                1    0    0    0    0
RAFACZ BA              4    0    0    0    0
RECRUIT                4    0    0    0    0
RELATION               2    0    0    0    0
SEELEY LC              3    0    0    0    0
SOLUTION               4    0    0    0    0
SR                     4    0    0    0    0
TEST                   3    0    0    0    0
TP                     3    0    0    0    0
TR                     1    0    0    0    0
USAF                   1    0    0    0    0
USN                    4    0    0    0    0
VOCATION               3    0    0    0    0
WARD JH                1    0    0    0    0
```

Figure 3. Inverted Keyterm File (IKØØØ4.).

program (BUILD), (2) a dictionary printing program (DICT), (3) a query-retrieval program (QUERY), and (4) a system maintenance program (CHANGE).[5]

QUERY PROGRAM

Introduction

The QUERY program represents the only component of the CARL System which will concern most users, since they will be searching a reference library which belongs to someone else. Those individuals who wish to set up their own computer-based reference library will need to become familiar with the entire CARL System of programs and data files. Since the primary design objective for the QUERY program has been simplicity of reference query and retrieval, effective use of the program will require no knowledge of computer programming.

Logical Search Operations

The simplest use of the QUERY program involves employing a single keyterm (keyword or author). However, the QUERY program allows more complicated requests, using combinations of keyterms. For example, it may be desirable to retrieve only those references which are jointly described by keyterms A and B. If a reference is characterized by keyterm A or keyterm B, but not by both keyterms, it is not considered a "hit". This conjunctive type of search is sometimes referred to as "AND-ing." Only those references described by keyterms A and B are desired.

Another type of multiple keyterm search allowed is disjunctive, and is sometimes referred to as "OR-ing". A disjunctive search allows the retrieval of all references which are described by either keyterm A or keyterm B (or both)

Finally, the QUERY program allows these two types of logical operations ("AND-ing" and "OR-ing") to be combined in a single search. For example, a search request can be made for all references which are jointly described by keyterms A, B, and C, plus a second set of all references described by either keyterm D or E. The "AND" list for input would include keyterms A, B, and C. The "OR" list would include keyterms D and E.

Search Modifications

After entering the keyterms for a request, the keyterms are printed out, and the user is given an opportunity to correct any errors. Then the inverted keyterm file is searched and the number of references located is printed out. At this point, the user can modify the original request with either type of logical operation. If the number of hits is too high, the user may wish to reduce the number of references by "AND-ing" one or more additional keyterms. Only those references described by the original keyterms and the additional keyterms are considered hits. A conjunctive type of modification will typically reduce the number of hits. On the other hand, the number of hits found for the

[5] The CHANGE program is in the design phase and is not operational at the present time.

original search may be smaller than desired and the user may wish to expand the search by "OR-ing" one or more additional keyterms. A disjunctive modification will typically increase the number of hits obtained.

Input Instructions

At the very beginning of an interactive session with the QUERY program, the user is informed that most questions require one of the following single character responses: "Y" - indicating yes; "N" - indicating no; "B" - indicating backup to the previous section; and "E" - indicating exit from the program. This response should be followed by pressing the carriage return to send the information to the computer.

When multiple keyterms are entered, they must be separated by commas and the entire list must be followed by a slash (/). Two separate lists are required for each multiple keyterm search: (1) an "AND" list and list. The maximum number of keyterms allowed for each of these lists is five. If the search request only requires an "OR" list, pressing the carriage return will bypass the "AND" list, and solicit the "OR" list. Similarly, if the request is a conjunctive type, then the "AND" list should be entered (followed by a slash). When the "OR" list is solicited, the carriage return should be pressed to bypass this input requirement.

As indicated above, a search can be modified to reduce the number of hits ("AND-ing"). The number of hits also can be expanded ("OR-ing"). When this necessity arises, the QUERY program explains the options in detail.

Output Instructions

Upon completion of each search query, the user is allowed to select both the extent and the location of the output. Four levels of output are available: (a) no output, (b) reference numbers, (c) reference citation, and (d) all information (includes the header card, the reference citation, the author card(s), and the keywords) for each hit. The user can direct the output to either the terminal or high speed line printer. The actual output options are presented to the user as follows:

Ø = No output
1 = Reference number(s) at terminal
2 = Reference number(s) to printer
3 = Reference listing at terminal
4 = Reference listing to printer
5 = All information at terminal
6 = All information to printer

The user enters the number corresponding to the option selected. If the user wishes to select more than one output option, this can be accomplished by selecting one option, then entering a "B" (for "backup") in response to the next question presented. This action returns the user to the list of output options where the second option desired can be entered. For example, there may have been 25 hits for a particular search. Due to the relatively slow printing speed of the terminal, the user might select option #4 to obtain a list of the reference citations for the 25 hits on the high speed printer. In addition, the user may wish to begin withdrawing the actual reference documents from the file cabinets housing the reference library, so the backup command is employed and then option #1 is selected. This action will cause the 25 reference numbers to be printed on the terminal for immediate use. The ability to direct the output to the high speed line printer also enables the user to proceed with another search without waiting for all the output to be produced on the terminal.

SAMPLE PROBLEMS

These sample problems are presented to illustrate the various types of search operations available to the user of the QUERY program. All examples are drawn from the previously mentioned sample reference library containing four references.

Single Keyterm Search

Suppose the user wishes to find all the references associated with the Armed Services Vocational Aptitude Battery. Examination of the user dictionary (Figure 1) shows that the appropriate keyterm for this search is "ASVAB." The inverted keyterm file (Figure 3) shows that only reference #3 has been characterized by the keyterm "ASVAB."

If the user wants to locate all the references written by Edward F. Alf, J the appropriate keyterm will be "ALF EF" and this inquiry will show that he is one of the authors of reference #2.

Conjunctive Keyterm Search

Suppose the user wishes to identify all references published by a military organization on the topic of recruiting. The user dictionary (Figure 1) indicates that the appropriate keyterms for this request are "MILPUB" and "RECRUIT." The inverted keyterm file (Figure 3) indicates that "MILPUB" is associated with references #1, 3, and 4, while "RECRUIT" characterizes reference #4. Since this inquiry is conjunctive, the number of hits will be one (i.e., reference #4). If the user wants only those references on recruiting that were published by military organizations during fiscal year 1976, the additional keyterm will be "FY76." Figure 3 shows that "FY76" is associated only with reference #2. The number of hits for this three keyterm conjunctive request will be zero, as no reference is jointly characterized by these three keyterms.

Disjunctive Keyterm Search

Suppose the user wishes to identify all publications written by either Edward F. Alf, Jr. or Joe H. Ward, Jr. The appropriate keyterms are "ALF EF" and "WARD JH." As indicated in Figure 3, "ALF EF" is associated with reference #2, while "WARD JH" is associated with reference #1. Since this request is disjunctive, the number of hits is two (i.e., references #1 and 2).

Mixed Keyterm Search

Suppose the user wishes to identify all military publications on the topic of recruiting. In addition, it is desired to locate all references authored by either Leonard C. Seeley or Joe H. Ward, Jr. The "AND" list for this request will be "MILPUB, RECRUIT/" and the "OR" list will be "SEELEY LC, WARD JH/." The conjunctive portion of this mixed keyterm query will produce one reference (i.e., reference #4). The disjunctive portion of the query will produce two references (i.e., references #1 and 3). The number of hits for this mixed keyterm query is three (i.e., references #1, 3, and 4).

CONCLUSIONS

The CARL System is a general and quite flexible document retrieval system. Unlike commercial reference retrieval systems, the source documents for the CARL System need not be published. This allows a particular CARL System to be dovetailed to the interests of a particular individual or organization. In the case of an individual researcher, the documents might include books, journal, newspaper and popular magazine articles, and lecture/class notes.

The management of an organization such as a military personnel research center could use the CARL System to store and retrieve information on all previous, on-going, and planned research projects. The source documents might include the DD-1498 Forms for the workunits and reports on milestones, funding, and endproducts.

The interactive QUERY program of the CARL System could be used to gather information quickly about a particular project or project area. This ability would expedite preparation of responses to the numerous internally and externally generated inquiries addressed to the organization management. In addition, all the visual aids (e.g., slides) which have been prepared for previous briefings, papers, etc. could be indexed by keyterms. Then, when a requirement for a new presentation arises, those visual aids pertinent to the topic could be retrieved quickly and efficiently.

While the CARL System can be used in any of the above applications, it is not meant to replace nor to compete with the various generalized information retrieval systems which are commercially available. These generalized systems accept input data into predefined fields without the special coding of keyterms which is necessary for the CARL System. The generalized systems also provide the capability of selecting and sorting information, and generating formatted reports. To make use of these valuable features, however, the user must know the names of the designated data fields and the relationships between these fields. In addition, the user of a generalized information retrieval system must learn a command language in order to formulate queries and direct system functions. The CARL System, on the other hand, requires no special knowledge of a command language. The user need only select the appropriate keyterms for a particular topic and respond to a choice of system options as they are presented. Therefore, anyone can make queries on the system with minimal preparation and little, if any, difficulty.

In summary, the CARL System appears to offer considerable promise for meeting many of the information storage and retrieval requirements of both researchers and managers.

REFERENCES

Broadhurst, P. L. Coordinate Indexing: A Bibliographic Aid. American Psychologist, 1962, 17, 137-142.

Lancaster, F. W. Information Retrieval Systems: Characteristics, Testing and Evaluation (2nd ed.). New York: John Wiley & Sons, 1979.

Sands, W. A. Computer Assisted Reference Locator (CARL) System: An Overview. Proceedings of the 20th Annual Military Testing Association Conference, October 1978.

THE EMBEDDED FIGURES TEST: ITS RELATIONSHIP TO FORWARD OBSERVER PERFORMANCE

Helen E. Belletti
Jack G. Anthony

Directorate of Lvaluation
US Army Field Artillery School, Ft. Sill, Oklahoma 73503

INTRODUCTION

A major concern of the Field Artillery School, Ft. Sill, Oklahoma is the training of the forward observer (FO). The function of the forward observer is to find hostile targets and report their locatins in an accurate and timely manner to a fire direction center. There, firing data is computed and directed to the firing batteries. If data is correct, the target is destroyed or disabled. If incorrect, adjustments are made and firing continues. One study estimated that as much as 50% of the error in the field artillery system can be attributed to the errors in target location and range reported by the forward observer.

In 1978 the College of Education, University of Oklahoma, conducted a study entitled "Evaluation of Forward Observers," (Officers) for the Directorate of Evaluation, USAFAS, in an effort to examine the relationship between forward observer performance and individual differences in learning ability. The main purpose was to investigate certain learner characteristics affecting forward observer performance. These learner characteristics were two cognitive styles (Visual-Haptic and Field Independence/Dependence), trait anxiety, and intelligence. One significant conclusion of the evaluation was that the cognitive style, "field independence/field dependence" was directly related to officer forward observer performance. It was found that field independent students tended to perform better as forward observers than did field dependent students. This characteristic was measured using the Embedded Figures Test, a commercially produced visual perception test. (See Methodology for further discussion of this test.) In light of these findings and the fact that training of 13F enlisted forward observers was inaugurated at Ft Sill, in March 1978 it was decided that an evaluation would be conducted by the Directorate of Evaluation, as a followup to the University of Oklahoma study.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

OBJECTIVES

To examine the relationship of field independence/field dependence to the performance of enlisted forward observers.

To determine the effectiveness of the Embedded Figures Test (EFT) as a vehicle for predicting enlisted forward observer success.

METHODOLOGY

The general design of the study followed the pattern usually used in empirical studies dealing with prediction problems (i.e., establishing the relationship between variables whose values are known prior to the onset of training with criterion measures gathered during or at the termination of training; if the relationships are of significant magnitude, then such information can be used in further iterations of the course either to select students whose probability of success is higher than a specified level or to modify the training program so as to insure a higher success rate or higher aggregate levels of achievement among the students that are selected).

Data used in this report are based on information gathered from 113 students in classes 6-79 through 11-79 of the four week Field Artillery Fire Support Specialist Course. This constituted the majority of students enrolled in the course during January through March 1979. (Of the 113 students, 87 graduated, 18 were relieved from the course, 3 were still attending the course at the time this report was written, and 5 had incomplete data.) Specific data on demographics, cognitive style, performance on the Armed Services Vocational Aptitude Battery and performance in self- and target-location were obtained with the instruments described in detail below.

The instrument for gathering demographic data was a student questionnaire which provided for the collection of information regarding student age, education level, prior military service and present military status, all of which have been found in previous studies to bear some relationship to course performance. Since handedness (right or left) and residency during the formative years (city, town or country) were suspected of being related to the type of performance expected of the forward observer, information on these variables was also requested in the questionnaire.

The instrument used in evaluating the cognitive style, "field independence/field dependence", was the Embedded Figures Test (Witkin, et al, 1971). The Embedded Figures Test is a visual perception test that measures the subject's ability to locate a previously identified simple geometric figure which has been embedded in a more complex figure. The test is timed and the scores represent the number of obscured simple figures the student can sucessfully identify out of a total of eighteen different, but separate trials. Students who score low on the test are identified as being "field dependent"; their perception is dominated by the whole field they view, whereas students scoring high on the test are identified as "field independent" students; they are better able to perceive discrete parts of the whole field they view. Since performance on the EFT has previously been used to predict performance on other perceptual tests and since the OU study (previously cited) demonstrated its relevance to successful performance in self- and target-location skills, the EFT was used in this evaluation to determine its usefulness as a predictor of success in a course designed to train enlisted forward observers. Aptitude data derived from the Armed Services Vocational Aptitude Battery (ASVAB) were collected directly from the students' personnel files. The two ASVAB scores used in this evaluation were the FA and GT scores. A minimum of 100 on the FA score is used as the selector for the course whereas the GT is a more general measure of ability which has been found in previous studies of this nature to have a relatively high relationship to academic performance in general.

Criteria of success used in this study were performance data involving both self- and target-location. Student performance levels were derived from data routinely collected during field exercises conducted by the Counterfire and Gunnery Departments. The measure of self-location performance was the difference between a student's position and his estimation of that position (radial missed distance, expressed in meters). This self-location score was taken from the Counterfire Department Map Reading exercise. The two criteria used for determining performance in target location were (1) radial missed distance averaged across four targets and (2) final shoot grade average achieved by the student in the four Gunnery Department Observed Fire exercises. The shoot grade is a broader measure of performance that takes into account proper procedure, number of rounds used, as well as the initial target location.

Frequency tables were constructed for each of the above mentioned variables. Relationships between the independent variables: education level, handedness, and resident location; and the dependent variables: radial missed distance score and average shoot grade, were investigated using Chi Square analysis. Relationships between the independent variables: FA score, GT score, and the EFT score; and the dependent variables: self-location radial missed distance score, target-location radial missed distance score, and average shoot grade were investigated using correlation analysis. Graduate and nongraduate performance on the EFT was computed, and attrition rates were calculated using an EFT cut-off score and the success/failure experience of this class. The expected effect of using a cut-off score was also computed as was the effect of using more than one predictor score. The latter was analyzed using a regression analysis of GT score, FA score, and EFT score.

FINDINGS

Findings in this report are presented in three categories; characteristics of the sample, performance of the sample, and results of statistical analysis performed on the above mentioned variables. Computations are based on data obtained from the 87 graduates of the course.

1. Description of Sample Characteristics:

TABLE 1. SAMPLE CHARACTERISTICS

AGE CATEGORY	PERCENTAGE*	N = 87
24 years and over	14.9	13
21 through 23 years	18.4	16
Less than 21 years	66.7	58
EDUCATION CATEGORY		
Some college education (or graduate)	6.8	6
High school graduate	55.2	48
Less than high school	37.9	33
SERVICE CATEGORY		
Prior service	9.3	8
No prior service	90.7	78
HANDEDNESS CATEGORY		
Right-handed	77.9	67
Left-handed	22.1	19
RESIDENT CATEGORY		
Country (rural)	15.1	13
Small town (100-4,999 residents)	19.8	17
Small city (5,000-99,999)	30.2	26
Metropolitan area (100,000+)	34.9	30

*Percentages in this report are rounded to the nearest tenth of one percent; therefore, cumulative percentages for each table may not always equal one hundred percent.

2. Description of Sample Performance:

TABLE 2. ARMED SERVICES VOCATIONAL APTITUDE BATTERY

SCORE INTERVAL	PERCENTAGE	
	FA Score	GT Score
131-150	4.6	3.4
121-130	10.3	9.2
111-120	27.6	28.7
101-110	49.4	33.3
91-100	5.7	11.5
70-90	2.3	12.6

Table 3 presents the percentages for the Embedded Figures Test scores. The scores represent the number correct out of a total possible score of eighteen. The data in this table represent scores of both graduates and nongraduates. Mean score for graduates is 9.4 and for nongraduates it is 5.7.

TABLE 3. DISTRIBUTION OF RESULTS OF EMBEDDED FIGURES TEST

SCORE INTERVAL	PERCENTAGE		
	Graduates (N = 87)	Academic Nongrade (N = 7)	Administrative Nongrade (N = 11)
16-18	16.1	0	9.1
13-15	18.4	0	0
10-12	12.6	0	18.2
7-9	18.4	14.3	18.2
4-6	20.7	42.8	36.4
0-3	13.8	42.8	18.2

Frequencies of self- and target-location scores are presented in Table 4. They represent radial missed distance raw scores where a score of "0" represents a perfect score. Self-location score is taken from one exercise and target-location score is an average taken from four exercises. The mean radial missed distance scores for each of the four exercises were: 457.6M, 537.5M, 428.7M and 363.8M. The range of scores for self-location is from 0 to 7100 meters and the range for target-location is from 154 to 1256 meters. Applying the ARTEP standards (250M for target-location and 150M for self-location), it was found that 82.8% of the students did not achieve ARTEP standard for target-location and 26.4% of the students did not achieve ARTEP standard for self-location.

TABLE 4. DISTRIBUTION OF SELF- AND TARGET-LOCATION SCORES

SCORE INTERVAL (Radial missed distance in meters)	PERCENTAGE* Self (N = 75)	 Target (N = 87)
1000 and over	10.4	2.2
900-999	0	0
800-899	0	3.3
700-799	0	3.3
600-699	2.7	6.6
500-599	1.3	12.1
400-499	2.6	12.1
300-399	5.4	33.0
200-299	8.0	18.7
100-199	45.3	4.4
1-99	0	0
0	24.0	0

Self-Location Mean Score = 368.3M (all cases)
134.1M (cases of 1000M+ excluded)
Target-Location Mean Score = 432.3M (all cases)
414.0M (cases of 1000M+ excluded)

Table 5 presents graduate percentages for shoot grade. This grade is an average taken from four field exercises. Mean grades for each of the four exercises were: 77.7, 83.3, 78.6 and 85.2. The grades range from 35% to 96%.

TABLE 5. AVERAGE SHOOT GRADE

Grade Interval	Percentage of Graduates
96-100	2.2
91-95	10.1
86-90	20.3
81-85	13.4
76-80	31.6
71-75	14.4
66-70	2.2
61-65	2.2
Below 60	1.1

3. Analysis of Data.

Neither the distribution of age nor that of prior military service sh::ws sufficient differentiation to warrant further analysis; therefore, the first three independent variables to be investigated for their possible effect on forward observer performance were education level, handedness, and resident location. Students were divided into two groups based ɔn two measures of performance; average shoot grade and average radial missed distance score for target location. Group #1 represents students scoring above the mean score in average shoot grade and below the mean in radial missed distance. (Recall that zero radial missed distance is perfect performance whereas 100% is perfect performance on the "shoot grade" criterion.) In each case, Group #1 represents those students displaying better than average performance. Group #2 represents students scoring below the mean for average shoot grade and above the mean in average radial missed distance. In each case, Group #2 represents students performing lower than average. Since the distribution on both criteria departs rather substantially from the normal distribution, the percentage of cases appearing in each of the two groups is not the same. Multiple frequency distribution tables were constructed for each variable and are presented in Tables 6, 7 and 8.

TABLE 6. RELATIONSHIP BETWEEN EDUCATION LEVEL AND PERFORMANCE

	CRITERION			
			Target-Location	
Years of Education	Grade Average		Radial Missed Dist. Average	
	(N = 87)		(N = 87)	
	Group 1	Group 2	Group 1	Group 2
13 - 16 years	N = 2 2%	N = 4 4%	N = 4 4%	N = 2 2%
12 years	N = 22 25%	N = 26 30%	N = 30 35%	N = 18 21%
9 - 11 years	N = 15 17%	N = 18 20%	N = 20 22%	N = 13 15%
	(chi square = 2.30) p = .89		(chi square = 4.32) p = .63	

The chi square values and their associated (p) values listed in Table 6 indicated that there is no significant relationship between education level and performance as a forward observer.

TABLE 7. RELATIONSHIP BETWEEN HANDEDNESS AND PERFORMANCE

Handedness	CRITERION Grade Average		Target-Location Radial Missed Dist. Average	
	Group 1	Group 2	Group 1	Group 2
Right-handed (N=67)	N = 30 35%	N = 37 43%	N = 44 51%	N = 23 27%
Left-handed (N=19)	N = 9 11%	N = 10 12%	N = 10 12%	N = 9 11%
(N=86)	(chi square = .004) (p = .95)		(chi square = .59) (p = .44)	

In Table 7, it can be seen that there is also no significant relationship between handedness and performance as a forward observer. (Chi square values are not significant.)

TABLE 8. RELATIONSHIP BETWEEN RESIDENCY AND PERFORMANCE

Resident Location	CRITERION Grade Average (N = 86)		Target-Location Radial Missed Dist. Average (N = 86)	
	Group 1	Group 2	Group 1	Group 2
Metropolitan area	N = 13 15%	N = 17 20%	N = 18 21%	N = 12 14%
Small city	N = 14 16%	N = 12 14%	N = 18 21%	N = 8 9%
Small town	N = 7 8%	N = 10 12%	N = 7 8%	N = 10 12%
Country (rural)	N = 5 6%	N = 8 9%	N = 11 13%	N = 2 2%
	(chi square = 1.17) (p = .75)		(chi square = 6.61) (p = .09)	

When resident location was addressed in this study to examine the relationship between a student's residence and his performance as a forward observer, it was theorized that students from a rural setting would perform better in target-location, especially in determiuing radial missed distance. In acting as a forward

Observer, students from a rural setting would be dealing with a task somewhat similar to the kinds of experiences provided by their rural environment and therefore, because of this familiarity, would be expected to perform better than a student from a more metropolitan area who had not been exposed to such experience. As can be seen in Table 8, this theory is partially substantiated in that they do perform better in target-location, but they do not perform better in grades earned. Further analysis shows that 85% of all students from a rural area scored above average on the radial missed distance criterion, whereas only 41% of students from a small town, 69% of students from a small city, and 60% of students from a metropolitan area performed in the above average group.

The next three variables to be examined in relationship to forward observer success were FA score, GT score and the Embedded Figures Test (EFT) score. At present, the FA score is used as a selector for the course. Each variable was correlated to determine its relationship to the three performance measures; self-location radial missed distance, target-location radial missed distance and average shoot grade. Presented in Table 9 are the results of this correlation analysis. The number of cases, the correlation coefficient (r) and the significance (s) is given for each correlation.

TABLE 9. RELATIONSHIP BETWEEN PERFORMANCE AND TEST SCORES (EFT, FA, GT)

Performance	FA Score	GT Score	EFT

Self-Location RMD	N = 75
	r = .016
	s = .45
		N = 74
		r = .114
		s = .17
			N = 75
			r = .015
			s = .45

Target Location RMD	N = 87
	r = -.18
	s = .05
		N = 86
		r = -.06
		s = .29
			N = 87
			r = -.28
			s = .005

Average Shoot Grade	N = 87
	r = .27
	s = .005
		N = 86
		r = .11
		s = .15
			N = 87
			r = .19
			s = .04

As seen in Table 9, there is a significant relationship between FA score and the two performance measures (target-location radial missed distance and average shoot grade) and the EFT score and these two performance measures. The GT score, however, shows no significant relationship to performance in any of the three cirteria used. It should be noted that the direction of the relationship for target-location is negative which is consistent with higher scores on the tests being associated with smaller error values in target-location.

After determining the correlations between the individual test scores (FA, GT, and EFT) and student performance (Table 9), a regression analysis was conducted to determine the effect of using the three scores together for a better prediction of target-location radial missed distance. (Results of the analysis are presented in Table 10.) It was found that using the three variable regression yields a multiple R of .32. For this set of data, this would mean that approximately 10% of the variation in target location score is accounted for by the variance in student performance on the FA, GT, and EFT. Although statistically significant, the relationship has little practical significance since it would only reduce the standard deviation in target location score by 9.5 points (from 208M to 198M). This suggests that the utility of such a procedure would not be sufficiently high to warrant its adoption.

TABLE 10. REGRESSION ANALYSIS: FA SCORE, GT SCORE, EFT SCORE

Component (Degrees of Freedom)	VALUE
Sum of Squares - Regression (3)	365363.29
Sum of Squares - Residual (82)	3230198.70
Mean Square - Regression (3)	121787.76
Mean Square - Residual (82)	39392.67
F	3.09
Multiple R	.32
R Square	.10

Since the Embedded Figures Test has shown to have some effect on forward observer performance, the anticipated effect of imposing a minimum score on the EFT as a course prerequisite was investigated. A minimum score of 7 was used, based on the fact that over 85% of the academic nongraduates achieved a score of 6 or less (see Table 3). This cut-off score was applied to the sample population used in this report and subsequent data was analyzed to determine the effect the use of this score would most likely have on input, number of graduates, number of failures and the resulting attrition rate. This is referred to in the table as the "constrained" condition. These data were then compared with data in which no such prerequisite was used (Actual). Table 11 summarizes these data.

TABLE 11. ANTICIPATED EFFECT OF IMPOSING EFT PREREQUISITE ON 13F COURSE

CATEGORY	NUMBER OF STUDENTS	
	Actual	Constrained (7+ on EFT)
Input	110*	67
Graduates	92**	61
Failures	18	6

ATTRITION RATE: Actual = 16%
Constrained = 9%

* Of the 113 original students, 3 were still enrolled at the time data were collected and were not included in this table.

** The 5 graduates who were not included in previous tables because of missing data are included here.

If the minimum score of 7 was to be used in selecting students for the 13F Course, there would be an anticipated reduction of 66% in the failure rate; that is, the number of students failing would be reduced from 18 to 6. However, the number of students entering the course would be reduced from 110 to 67, and the number of graduates would be reduced from 92 to 61.

CONCLUSIONS.

Students perform better in self-location than they do in target-location. (This is evidenced in that 83% of students do not achieve ARTEP standards for target-location, whereas only 26% do not achieve ARTEP standards for self-location.)

Age, education level and handedness do not significantly affect performance as a forward observer.

Resident location from six to twelve years of age does appear to have some relationship to a student's performance as a forward observer, especially his ability to determine target-location radial missed distance. Approximately 85% of the students from a rural setting scored above average in determining target-location radial missed distance.

Field independent students were found to perform better as forward observers than field dependent students; however, the Embedded Figures Test, used to test field independence/field dependence, taken alone was not found to predict forward observer success to a much greater degree than the present predictor, FA score.

Although the combination of GT score, FA score and the EFT score bears a significant relationship to success, its use would not yield any practical advantages. Estimates derived from this study suggest that if all three tests were to be used as predictors the error in predicting target-location radial missed distance average would only be diminished by about 10 meters over what it would be without knowing how well students perform on the various tests. This reduction in error is too small to warrant serious consideration.

Imposing an EFT minimum score as a prerequisite for the course would reduce the failure rate, but would also significantly reduce the number of students entering the course who would have graduated even though they did not meet the EFT criteria. In a context in which the 13F MOS is seriously undersubscribed, further constriction of the pool of eligibles, by imposing still another prerequisite, would be inappropriate at this time.

PSYCHOLOGICAL PROFILES AND PERFORMANCE CHARACTERISTICS OF TACTICAL SIMULATOR CONSOLE OPERATORS

BY

Charles W. Howard, Ph.D.
Us Army Research Institute for the Behavioral and Social Sciences
Ft. Bliss, Texas

ABSTRACT

An analysis of psychological profiles and performance data can provide training developers with guidelines for selection and assignment of tactical console operators. This paper presents descriptive performance data for console operators collected on a tactical simulator/trainer. Psychological data to include personality factors, general intelligence, numerical ability, visual pursuit, reading level, visual speed and accuracy are also reported. Relationships between performance data and psychological profiles are presented and discussed.

INTRODUCTION

The United States Army has established two personnel management systems. These systems are (1) Officer Personnel Management System and (2) Enlisted Personnel Management System. A keystone to each of the systems is training. Training as a subset of each of these systems has two elements. They are recruitment and placement. Recruiting candidates for available and projected jobs in an organization requires a knowledge of the expectations for the individual. Given a set of job demands, the organization is able to match capabilities of individuals with the job demands or requirements. An organization can enhance its level of productivity and its level of effectiveness when prerequisite cognitive processing, affective and psychomotor requirements are identified for a given set of tasks.

Modern air defense weapon systems are becoming more complex and potentially more demanding for both officer and enlisted personnel. The research to be reported was designed to assist the training designers and training developers in identifying the prerequisite characteristics desirable trainees should possess prior to entry into a job. Included in the project was the development of console operator data base encompassing both psychological and performance data.

One objective of this research project was to develop one or a set of predictive models that would enable the Army to improve their ability to predict success of a given trainee for a given set of tasks. The model would form a basis for developing screening instruments to maximize the percent of students successfully completing a given course of instruction. Several benefits in developing a console operator data base for recruitment and placement have been noted (Howard, 1978).

METHOD

Subjects

The sample consisted of twenty-eight United States Army Air Defense soldiers. Each of the subjects was assigned within the Fort Bliss Air Defense Center. The sample consisted of eleven officers and seventeen enlisted personnel. The organizations tasked to provide support for this research project were given two guidelines on the prerequisites of soldiers for this project. The guidelines primarily consisted of (1) required time for availability and (2) security classification.

Materials

The research project consisted of three major phases, a training phase, a hands-on testing phase, and a psychological testing phase.

The training phase consisted of a set of performance-paced materials (Howard and Hoffer, 1978). The training materials were designed to familiarize the subjects with nomenclature, location and operation of a tactical operations simulator/trainer (see Figure 1). The training session was conducted on either Tuesday or Thursday of a given week. The subjects began the session at approximately 0800 on a given day and completed the session in an average of four hours. Instructors were available during the training session to monitor student activities. At the completion of the training session a subject was given a written performance test (WPT-1) to evaluate the acquisition of his/her knowledge of the training objectives designed for the research project. Subjects not achieving mastery of the training objectives were given remedial instruction on the objectives not mastered. Subsequent training was then provided on procedures and skills with specific time critieria. (Howard, 1979) At the completion of this training and evaluation, subjects were instructed to take a second written performance test (WRT-II).

The hands-on testing phase consisted of approximately five hours of testing on the TOS/T. The psychological and psychomotor testing phase included the following set of test instruments: Employee Aptitude Survey Tests; Numeric Ability, Visual Pursuit, Space Visualization, and Numerical Reasoning; Otis Quick-Scoring Mental Ability Test: New Edition, Gama Test: Form FM; Cattell's sixteen personality factors, Form C; Lateral Perception Span (USARI); Gates-MacGinite Reading Test, Survey E; Pursuit Rotor Test with ten trials; Witkins - Group Embedded Figure Test, in addition to other tests not reported in this paper. Subjects were also given a Human Factors Evaluation Questionnaire and interviewed on the areas of training, environment, and equipment (Hoffer and Howard, 1979).

Procedure

A flowchart of the five phases is shown in Figure 2. The phases included:

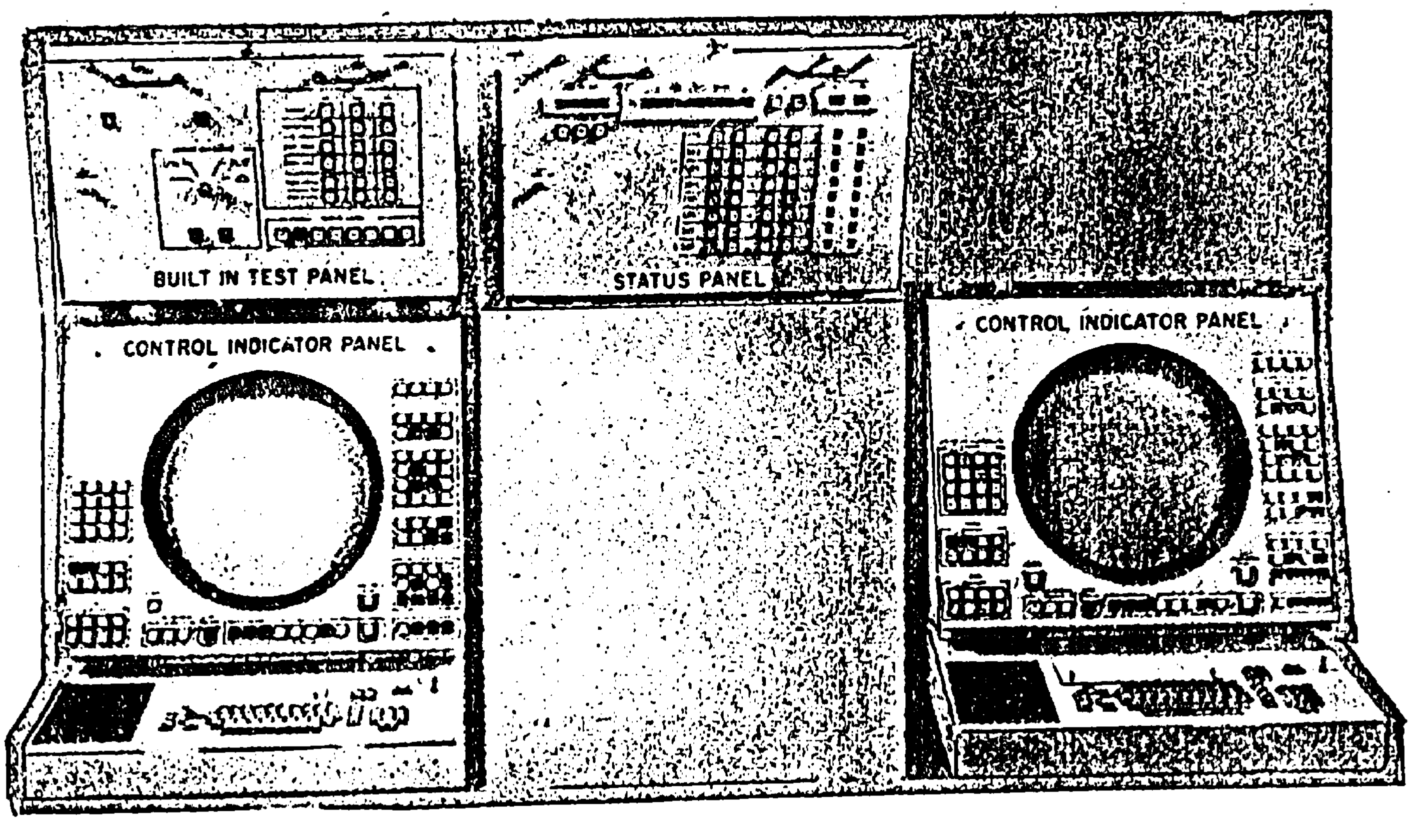


Figure 1. Operator console in Patriot Weapon System

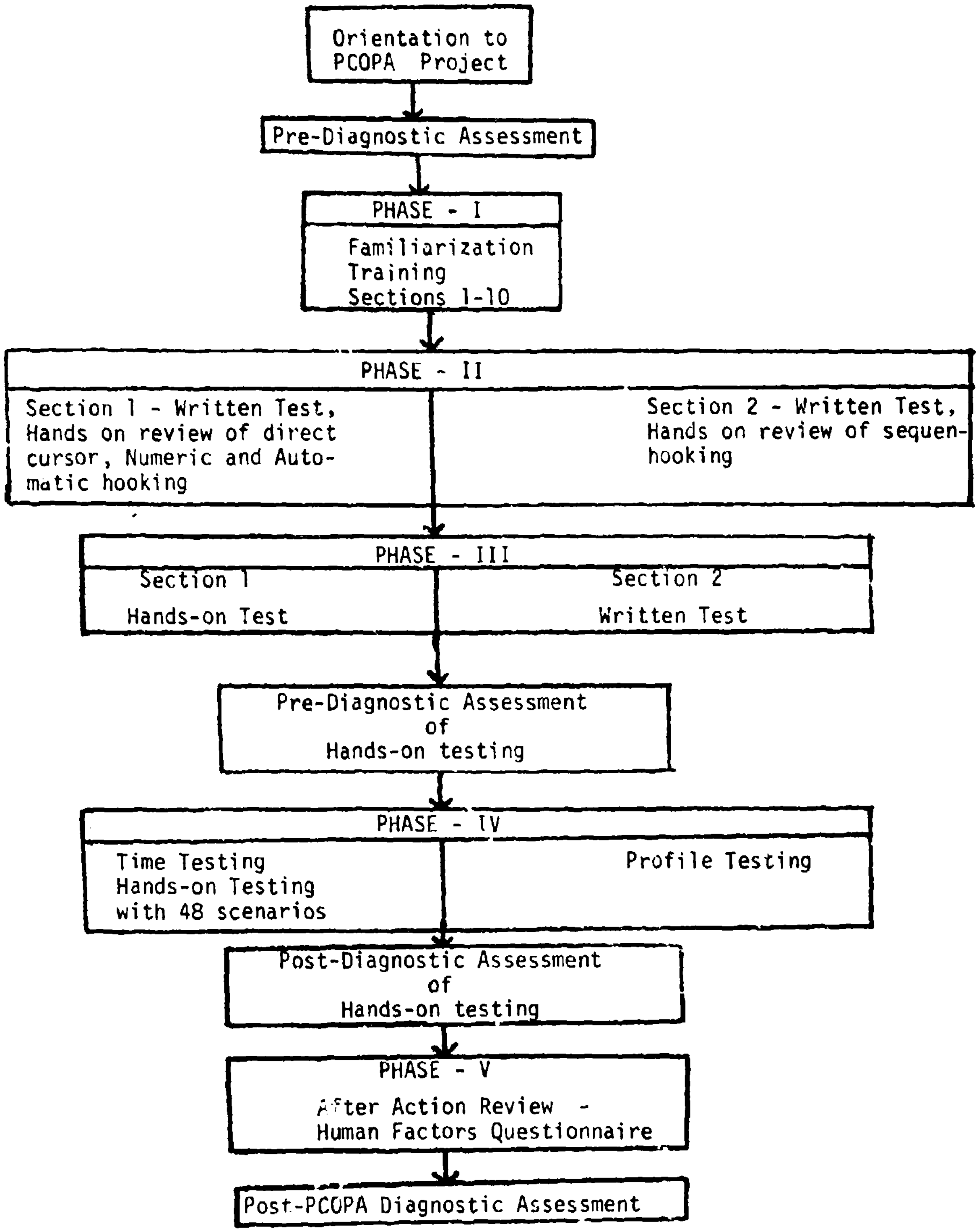

Orientation to PCOPA Project
Pre-Diagnostic Assessment
PHASE - I
Familiarization Training Sections 1-10
PHASE - II
Section 1 - Written Test, Hands on review of direct cursor, Numeric and Automatic hooking
Section 2 - Written Test, Hands on review of sequen- hooking
PHASE - III
Section 1
Hands-on Test
Section 2
Written Test
Pre-Diagnostic Assessment of Hands-on testing
PHASE - IV
Time Testing Hands-on Testing with 48 scenarios
Profile Testing
Post-Diagnostic Assessment of Hands-on testing
PHASE - V
After Action Review - Human Factors Questionnaire
Post-PCOPA Diagnostic Assessment

Figure 2. Student Activities for PATRIOT Console Operator Performance Analysis (PCOPA)

PHASE 1 - Familiarization Training (approximately four hours with all students working in a team training environment). During this phase the students were given the performance-paced training materials and scenarios to familiarize the students with the location and operation of the console buttons and keys. In addition, the students are taught how to perform basic tasks such as activating the console, hooking targets, and engaging targets.

PHASE 2 - Practice (approximately 45 minutes per student). Each student was given time on the TOS/T to practice the procedures taught in Phase 1.

PHASE 3 - Evaluation (approximately 45 minutes per student). Once the student had acquired the skills and knowledge to perform the basic operator tasks, he was evaluated for a set of tasks using a criterion-based evaluation checklist. If the student fails the evaluation, more practice was provided. (Often there was not enough time for additional practice because of the limited time available on the TOS/T. In these instances the instructor verbally explained the procedures and helped the student correct his incorrect responses. After this additional instruction was provided, the instructor assesses the student's ability to perform console operations and either dismisses the student from the program or allows him to continue.)

PHASE 4 - Testing (approximately five hours per student for hands-on testing and four hours per student for diagnostic testing). As noted earlier, each student was tested on the 48 standard scenarios. The performance data collected include each operator's button and keyboard responses and the time associated with each action for all hands-on testing scenarios. Data was also being collected on detection and intercept times, etc., for each operator for all scenarios. In addition, each student was given a battery of diagnostic tests to measure the cognitive and affective learning styles.

PHASE 5 After Action Review (AAR) (approximately two hours per student). After all students for a particular week had completed the research program they were interviewed. The interview questionnaire attempted to gather information from each student about the training, equipment, and environment.

The research program is an ongoing research effort. Presently, twenty-eight students have completed the program as either validation or test subjects. An anticipated 60 to 75 test subjects will complete the program before its conclusion of data collection in November 1979.

RESULTS

Means and standard deviation of the psychological, psychomotor, and written performance tests are shown in Table 1. Table 2 presents the correlations of the fifteen variables. Subsequent analysis of the data included stepwise multiple regressions. This analysis permitted the researcher to examine the best linear prediction equation and evaluate its prediction accuracy for the following dependent variables: written test - I, written test - II, and the sum of written test - I and written test - II using three regression models.

Table 3 presents the summary of the stepwise regression analysis (Model-1) with the criterion variable written performance test - I and ten predictors.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR THE PSYCHOLOGICAL AND PERFORMANCE VARIABLES

# VARIABLE	n	$\overline{X}$	(SD)
(1) Group Embedded Figure Test	28	11.39	5.45
(2) E.A.S. - Numerical Ability	28	38.03	18.26
(3) E.A.S. - Visual Pursuit	28	15.42	4.70
(4) E.A.S. - Space Visualization	28	27.53	10.34
(5) E.A.S. - Numerical Reasoning	28	10.10	4.62
(6) Lateral Perception Span (ARI)	28	37.35	7.31
(7) Gates - MacGinn. Vocab.	28	11.58	2.16
(8) Gates - MacGinn. Comp.	28	12.07	1.57
(9) Cattell 16PF, Factor B	28	4.64	1.87
(10) Cattell 16 PF, Factor I	28	4.64	1.59
(11) OTIS	28	109.32	15.12
(12) Pursuit Rotor	28	6.54	1.99
(13) Written Test - I	28	20.75	4.91
(14) Written Test - II	28	24.46	5.53
(15) Written Test - I & II	28	45.21	9.93

TABLE 2

CORRELATION MATRIX FOR PSYCHOLOGICAL AND PERFORMANCE VARIABLES 1/

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Group Embedded Figure Test	(1)	-	73	65	84	79	63	61	64	65	25	80	41	64	84	78
E.A.S.-Numerical Ability	(2)	*	-	37	66	72	48	58	46	80	44	87	14	58	75	70
E.A.S.-Visual Pursuit	(3)	*	03	-	73	47	60	51	29	12	55	50	50	54	66	64
E.A.S.-Space Visualization	(4)	8	*	*	-	67	60	59	71	67	29	88	47	58	79	73
E.A.S.-Numerical Reasoning	(5)	*	*	*	*	-	55	37	74	40	40	73	26	62	70	70
Lateral perception Span (ARI)	(6)	*	*	*	*	*	-	27	31	40	02	50	37	44	60	56
Gates-MacGinn. Vocab.	(7)	*	*	*	*	03	08	-	83	45	20	71	20	63	68	69
Gates-MacGinn. Comp.	(8)	*	*	06	*	*	05	*	-	42	24	61	30	62	77	73
Cattell 16 PF, Factor B	(9)	*	*	2,	*	*	02	*	*	-	61	78	20	48	63	58
Cattell 16 PF, Factor I	(10)	09	*	*	07	*	45	16	11	*	-	35	-17	16	38	29
OTIS	(11)	*	*	*	*	*	*	*	*	*	03	-	32	66	81	78
Pursuit Rotor	(12)	02	24	*	*	09	03	16	06	15	03	05	-	36	37	38
Written Test - I	(13)	*	*	*	*	*	*	*	*	*	21	*	03	-	81	94
Written Test - II	(14)	*	*	*	*	*	*	*	*	*	02	*	03	*	-	95
Written Test - I & II	(15)	*	*	*	*	*	*	*	*	*	*	*	02	*	*	-

1/ Correlations are presented on the upper triangular, one-tail test of significant values are presented in the lower triangular. Note: All values are rounded to the hundredths and decimal points are omitted.

* - P≤.01

TABLE 3

SUMMARY OF STEPWISE REGRESSION FOR WRITTEN PERFORMANCE TEST-I 1, 2/

VARIABLES IN EQUATION					VARIABLES NOT IN EQUATION		
Step	Variable Entered	R	df	F 3/	Variable	Partial Correlation	F to Enter
1	11	.663	1/26	20.43	1	.248	1.60
					2	-.093	.222
					3	.288	2.25
					4	.068	.117
					5	.267	1.93
					6	.170	.753
					7	.299	2.46
					8	.348	3.44
					9	-.083	.174
2	11 8	.719	2/25	13.11	1	.150	.550
					2	.193	.970
					3	.213	1.14
					4	-.120	.353
					5	.297	2.32
					6	.178	.790
					7	.069	.114
					9	-.132	.433
3	11 8 5	.746	3/24	9.97	1	-.156	.582
					2	-.082	.156
					3	.185	.815
					4	-.196	.916
					6	.096	.213
					7	.224	1.22
					9	-.181	.776
4	11 8 5 7	.760	4/23	7.85	1	-.080	.143
					2	-.122	.330
					3	.243	1.39
					4	-.091	.184
					6	.122	.331
					9	-.213	1.04
5	11 8 5 7 3	.776	5/22	6.67	1	-.128	.352
					2	.190	.791
					4	-.259	1.51
					6	.171	.618
					9	-.151	.491
6	11 8 5 7 3 4	.793	6/21	5.94	1	-.030	.189
					2	.305	.446
					6	.075	.114
					9	-.241	.759
7	11 8 5 7 3 4 6	.794	7/20	4.89	1	-.046	.396
					2	-.009	.186
					9	-.380	.631
8	11 8 5 7 3 4 6 9	.796	8/19	4.09	1	-.087	.394
					2	.009	.152
9	11 8 5 7 3 4 6 9 1	.796	9/18	3.46	2	.383	.108
10	11 8 5 7 3 4 6 9 1 2	.796	10/17	2.94			

NOTES: 1. See variable names of independent or predictor variables in Table 2.

2. Note: Variables listed in Table 2 consist of all variables used in the three regression models.

3. The F ratio for the overall R at each step.

The multiple R=.796 with an R squared=.633 and an adjusted R squared=.418 with an overall F significant at $p<.02$ was obtained on step ten of the stepwise regression analysis. The independent variables, pursuit notor and Cattell 16PF-Factor I were not included in this model due to the correlations with the written performance test - I.

Table 4 presents the summary of the stepwise regression analysis (Model-II) with the criterion variable written performance test - II and twelve predictors. The multiple R=.957 with an R squared=.916 and an adjusted r squared=.850 with an overall F significant at $p<.001$ was obtained on step twelve of the stepwise regression analysis.

Table 5 presents the summary of the stepwise regression analysis (Model-III) with the criterion variable written performance tests-I and II total score (additive sum) and twelve predictors. The multiple R=.909 with an R squared= .828 and an adjusted R squared=.709 with an overall F significant at $p<.001$ was obtained on step eleven of the stepwise regression analysis.

DISCUSSION

The findings indicate there is a significant linear relationship between the set of predictors and criterion variable for each of the three models. The author has planned for this paper to present a descriptive view of the data collected thus far in the research project. The sufficiency of the progammatic approach to achieving a refined recruitment procedure appears to be appropriate given the multiple correllations obtained. Further exploration of the utility of the refined recruitment procedure will include the ability to predict console operator hands-on performance. For this reason, the regression equations are not presented in this paper. The future identification of regression equations will include two major factors: (1) utility and (2) efficiency. Utility will consist of the potential users ability to implement the model with a known level of certainity. Efficiency will be governed by the trade off analysis of test administration "time" and accuracy of hands-on performance for the tactical operations simulator/trainer. The ability to predict written performance is only one of the steps in developing a feasible recruitment model for use by training designers and trainer developers with the complex time dependent tasks presented to operators of modern weapon systems of the future. As noted in Model II and III the predictor variable (pursuit rotor) or psychomotor variable had a low partial correlation with the written performance test-II and the sum of written performance tests I and II. The discriminate power for predicting acceptable hands-on performance and predicting hands-on performance that would not be acceptable will be the next step of this research. Future research will thus include the review and analysis of the discriminate functions using hands-on performance as a criterion. This procedure for generating discriminate functions may include multiple criterion variables when the reaction time data and effectiveness measures are available. Therefore, a secondary task will be to apply canonical correlation on the set of criterion variables and set of predictor variables. This technique and its relationship to multiple regression and discriminate analysis is described in the literature (Howard, 1976).

TABLE 4

SUMMARY OF STEPWISE REGRESSION FOR WRITTEN PERFORMANCE TEST - II 1,2/

VARIABLES IN EQUATION					VARIABLES NOT IN EQUATION		
Step	Variable Entered	R	df	F3	Variable	Partial Correlation	F to enter
1	1	.899	1/26	108.94	2 3 4 5 6 7 8 9 10 11 12	.394 .291 .282 .119 .170 .387 .561 .191 .315 .428 .044	3.99 2.32 2.17 .36 .82 4.41 11.49 .97 2.76 5.56 .99
2	1 8	.893	2/25	49.34	2 3 4 5 6 7 9 10 11 12	.428 .241 .086 .253 .317 -.032 .232 .311 .374 .026	5.41 1.48 .76 1.64 2.49 .28 1.37 2.57 3.90 .144
3	1 8 2	.913	3/24	40.54	3 4 5 6 7 9 10 11 12	.346 -.012 .138 .336 -.229 -.049 .170 .136 .137	3.36 .44 .39 2.93 1.28 .54 .69 .28 .58
4	1 8 2 3	.925	4/23	34.55	4 5 6 7 9 10 11 12	-.228 .164 .232 -.214 .001 .183 .015 .052	1.19 .67 1.25 1.54 .23 .76 .94 .60
5	1 8 2 3 6	.930	5/22	28.20	4 5 7 9 10 11 12	-.235 .128 -.195 -.001 .244 .038 .031	1.47 .35 .83 .51 1.33 .30 .27
6	1 8 2 3 6 4	.934	6/21	24.25	5 7 9 10 11 12	.150 -.2[illegible] .124 .274 .170 .095	.46 1.62 .32 1.63 .66 .78
7	1 8 2 3 6 4 7	.940	7/20	21.66	5 9 10 11 12	.080 .174 .224 .428 .118	.70 .99 1.01 4.21 .33
8	1 8 2 3 6 4 7 11	.951	8/19	22.74	5 9 10 12	-.118 .163 .250 .064	.25 .49 1.20 .79
9	1 8 2 3 6 4 7 11 10	.954	9/18	20.56	5 9 12	-.193 .019 .162	.64 .38 .35
10	1 8 2 3 6 4 7 11 10 5	.956	10/17	18.21	9 12	.109 .164	.195 .329
11	1 8 2 3 6 4 7 11 10 5 12	.957	11/16	15.94	9	.070	.75
12	1 8 2 3 6 4 7 11 10 5 12 9	.957	12/15	13.17			

NOTES 1. See variable names of independent or predictor variables in Table 2.

2 Note Variables listed in Table 2 consist of all variables used in the three regression models

3 The F ratio for the overall R at each step

TABLE 5

SUMMARY OF STEPWISE REGRESSION FOR WRITTEN PERFORMANCE TEST - I & II 1, 2/

VARIABLES IN EQUATION					VARIABLES NOT IN EQUATION		
Step	Variable Entered	R	df	F3	Variable	Partial Correlation	F to Enter
1	1	.784	1/26	41.53	2 3 4 5 6 7 8 9 10 11 12	.290 .277 .200 .202 .130 .43[illegible] .490 .190 .182 .410 .107	2.43 2.09 1.07 1.12 .42 5.73 7.93 .85 .89 5.06 .79
2	1 8	.841	2/25	30.15	2 3 4 5 6 7 9 10 11 12	.344 .226 -.099 .320 .231 .117 .101 .115 .300 .086	3.13 1.30 .21 2.93 1.36 .39 .82 .32 2.46 .22
3	1 8 11	.862	3/24	23.31	2 3 4 5 6 7 9 10 12	.130 .230 -.208 .246 .226 -.040 -.032 .030 .108	.45 1.88 1.02 1.49 1.35 .35 .24 .21 .27
4	1 8 11 5	.871	4/23	18.21	2 3 4 6 7 9 10 12	.103 .281 -.201 .225 .050 -.143 -.048 .130	.23 1.81 .03 1.17 .58 .47 .53 .64
5	1 8 11 5 3	.881	5/22	15.28	2 4 6 7 9 10 12	.193 -.379 .144 .130 -.074 -.033 .056	.85 3.43 .47 .31 .12 .34 .72
6	1 8 11 5 3 4	.890	6/21	14.72	2 6 7 9 10 12	.155 .207 -.124 .043 -.010 .140	.49 .90 .32 .41 .66 .43
7	1 8 11 5 3 4 6	.903	7/20	12.69	2 7 9 10 12	.130 -.134 .042 .026 .135	.32 .36 .43 .14 .35
8	1 8 11 5 3 4 6 7	.905	8/19	1[illegible].78	2 9 10 12	.124 .092 .018 .128	.29 .16 .64 .30
9	1 8 11 5 3 4 6 7 12	.906	9/19	9.27	2 9 10	.144 .094 .043	.45 .15 .30
10	1 8 11 5 3 4 6 7 12 2	.909	10/17	8.19	9 10	.035 -.019	.20 .60
11	1 8 11 5 3 4 6 7 12 2 9	.909	11/16	6.79	10	-.006	.21

NOTES 1. See variable names of independent or predictor variables in Table 2.

2. Note. Variables listed in Table 2 consist of all variables used in the three regression models.

3 The F ratio for the overall R at each step

REFERENCES

Hoffer, P.L., Howard, Charles W., Human Factors Evaluation of an Air Defense Training Simulator - TOS/T. Us Army Research Institute for the Behavioral and Social Sciences, Ft Bliss, TX. Working Paper (submitted for publication) Sept 1979.

Howard, Charles W. "A Review of the Multivariate Technique Canonical Correlation," Journal of the Institute for Multidisciplinary Graduate Research. Washington, D.C: The Catholic University of America, 1:1, May 1976.

Howard, Charles W., Methodology for Evaluating Operator Performance On Tactical Operational Simulator/Trainers, US Army Research Institute for the Behavioral and Social Sciences, Ft Bliss, TX, a Paper pressented at the 20th Annual Conference of the Military Testing Association, October 1978.

Howard, Charles W., Hoffer, Peggy L., PATRIOT Man-Machine Interface: A Cooperative Research Analysis, US Army Research Institute for the Behavioral and Social Sciences, Ft Bliss, TX, Familiarization Manual for TOS/T, Research Product, 1978.

Howard, Charles W., Criterion Based Evaluation Procedure for PATRIOT Console Operator Performance Analysis, US Army Research Institute for the Behavioral and Social Sciences, Ft Bliss, TX, Working Paper, 1979.

AUTHOR

HOWARD, CHARLES W. Address: US Army Research Institute for the Behavioral and Social Sciences, Ft Bliss, TX 79916. Title: Research Statistician. Degrees: B.S., University of Charleston, West Virginia; M.A., Ph.D., The Catholic University of America. Specialization: Educational Technology, Applied Statistics, Man-Machine Interface Analysis. Telephone: AV 978-4491, Commercial (915) 568-4491.

A FORWARD OBSERVER QUESTIONNAIRE: PERSONNEL SELECTION AND TRAINING IMPLICATIONS[1]

John B. Mocharnuk and Dawn S. Trelz

McDonnell Douglas Astronautics Company
Saint Louis, Missouri

Raymond O. Waldkoetter

U.S. Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, Fort Sill, Oklahoma

The test development activity described here was one component of a larger research effort designed to identify ways of improving Field Artillery Forward Observer (FO) performance and to document the basic FO job. The former requirement resulted from an awareness that FOs were performing below Army Training and Evaluation Program (ARTEP) standards. The latter requirement was necessitated by a changing environment brought about by the implementation of the Fire Support Team (FIST) concept. A questionnaire was developed to identify effective FOs and to serve as an aid in describing learner characteristics. The development of that questionnaire, creation of models of observed fire performance, and potential applications of the questionnaire are described below.

Development of the Forward Observer Personal Profile Questionnaire. In order to identify background factors, interests, activities, and abilities of the FO population, Developmental Form A of the Forward Observer Personal Profile Questionnaire (FOPPQ) was developed as a research instrument. In developing the questionnaire an attempt was made to find items which would provide a broad distribution of scores and, additionally, to identify those factors which could be expected to have a relationship to individual combat effectiveness as measured by several criteria.

A device developed specifically to differentiate among combat effective and combat ineffective pilots, the Pilot Life Inventory Questionnaire (Youngling, Levine, Mocharnuk, and Weston, 1977), and results of an item analysis on that device proved useful in developing the FOPPQ. Some items from the Pilot Life Inventory Questionnaire were adapted for use in the FOPPQ, and a few others were from a test which has related attitudinal, achievement, and personal data to on the job performance (Nelson, Marco, and Banks, 1976) and the Division 14 (of the American Psychological Association) file of personal information items.

The FOPPQ was divided into five sections, A through E, according to the type of answers required and the general nature of the data sought. The former division was for ease of administration and data reduction; whereas, the latter was to allow selective use of sections, if necessary, and to provide continuity for those completing the questionnaire. This test format additionally provided an organizational scheme for potential users of the questionnaire data. Section A consisted of life experience and activities questions where multiple responses could have been appropriate. Items in Section A included varied topics from participation in sports to use of calculators. Section B included life experience items which required a single response. Topics included size of hometown and identification of courses in which the student received the highest grades. Section C also required a single response but focused specifically on issues pertaining to being in the Army. Sections D and E included attitude questions with responses from strongly agree to strongly disagree. Section D focused on FO related issues, and Section E had a broader scope.

Development of Preliminary Selection Models

Preliminary selection models were developed using the same categories of personal information which were used in the description of the FAOBC population. The first step in consideration of selection models involved an assessment of criteria. The second step was a modeling activity which yielded predictive models of FO performance.

[1] This is based upon research conducted for the U.S. Army Research Institute for the Behavioral and Social Sciences under contract DAHC-19-78-C-0025 which is reported in a report entitled "Selection and Training of Field Artillery Forward Observers: Methodologies for Improving Target Acquisition Skills."

Criterion Selection. The criteria, or measures of performance against which selection devices are evaluated, were selected according to pragmatic considerations. They had to be both reasonable and accessible. Three criteria were ultimately selected and received varying degrees of analysis. The first was target location accuracy. This was defined as first round target location error or radial miss distance (RMD), the distance from the student's specification of target location and actual target location, for selected shoots. The second was the GØ-0211 observed fire score. The GØ-0211 score is an observed fire grade based upon all graded firing exercises for the student and the best two of three hasty target location exercises. In determining the grade for each shoot, the instructor must include some subjective elements such as relative target location difficulty, but the grades are based primarily upon the accuracy, speed, and procedural adequacy with which the mission was handled. The third was overall success in FAOBC as reflected in the final grade.

Unfortunately other potential criteria from operational units were not available. First, standardized measures of FO performance do not exist. Second, even if the measures existed, no adequate means exist for tracking and extracting detailed measures of individual performance beyond the U.S. Army Field Artillery School environment without infringement on individual privacy. Measures of performance in the operational environment are necessary if a test is to be validated against intermediate operational criteria rather than school based measures. These difficulties did not preclude an effective selection of criterion measures. Instead they forced a more thorough analysis of the potential criterion measure which could be recorded and used. The criterion set was restricted to intermediate criteria which included training and performance measures collected at Fort Sill. The performance-based measures, RMD and the Observed Fire Score, GØ-0211, emerged as candidate criterie early and were of special interest because they could be expected to be more directly related to measures taken in unit testing environments, e.g., ARTEPS, than paper and pencil test scores.

Another step in the criterion analysis process consisted of an evaluation of the interrelationships of severel of the potential criterion measures. A spanning tree and a hierarchical tree were developed from a correlation matrix of selected criterion measures. The measures selected for evaluation were pinpointed through discussions with instructors as most relevant to FO task performance. Table 1 includes a list of FAOBC grades with a brief description of the domain covered by each one. Weapons Department scores were excluded from consideration as measures of FO performance. Those measures reflected an area of the artillery officer's responsibility outside of the FO job. The correlatiom matrix of scores recorded for the 175 FAODC Sample I students for whom final FAOBC grades were available was used in this analysis. In developing the spanning tree a nearest neighber algorithm was used. That is, the spanning tree was built by first selecting those two scores with the highest correlation as reflected in the correlation matrix of eleven selected FAOBC grades. At the second step the score which had the highest correlation with either of those two scores was selected. Next, the item with the highest correlation with any of the previously selected scores was selected and so on. The lines connecting the spanning tree represent an organization along the strongest connections.

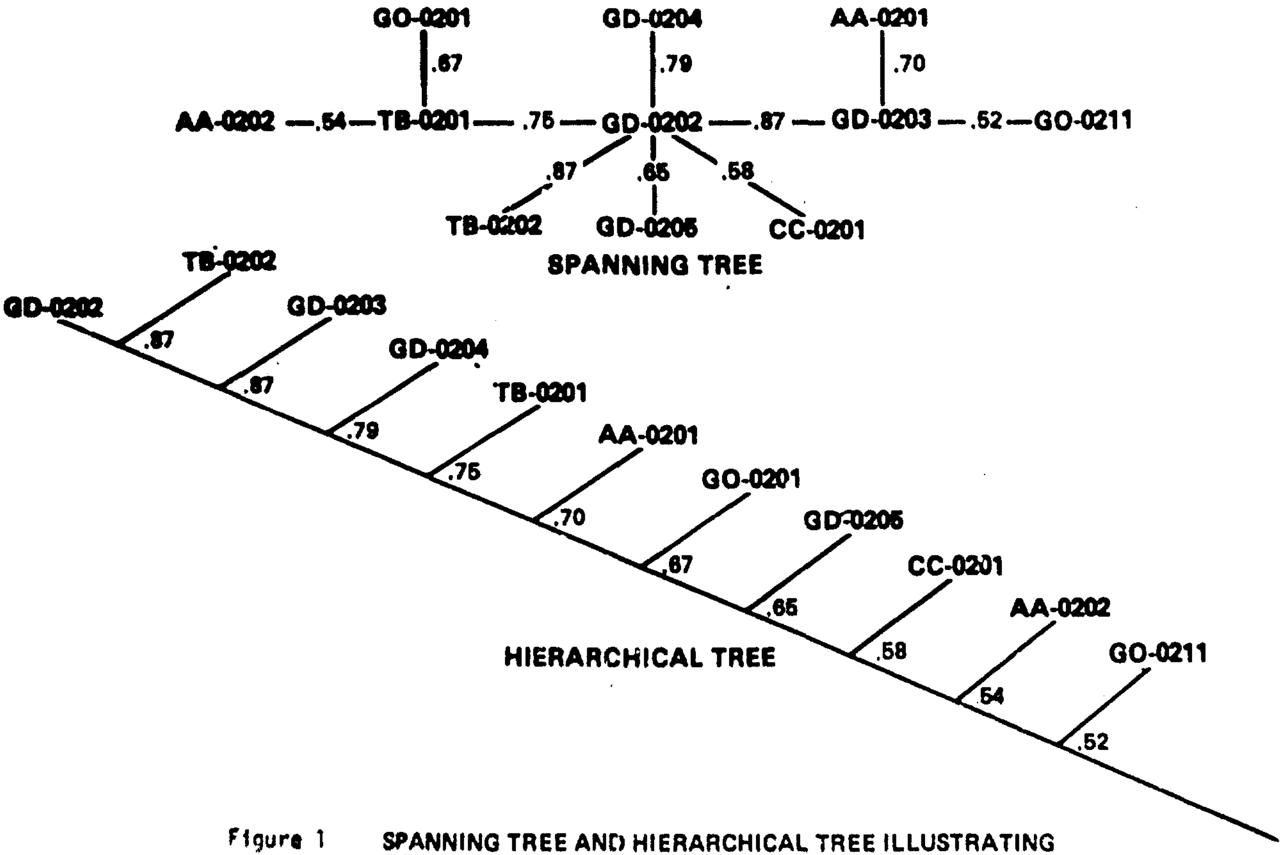
Table 1

OBC Component Grades

Grade	Dept.	Weight	Description
AA-0201	CFD	38	Map Reading Practice
AA-0202	CFD	32	Targeting
CC-0201	C&E	20	Communications
CC-0202	C&E	30	Communications
GD-0202	GD	80	Fire Direction
GD-0203	GD	70	Fire Direction
GD-0204	GD	25	FADAC
GØ-0201	GD	50	Observed Fire Written
GØ-0211	GD	125	Observed Fire Practical
TB-0201	TCAD	90	Artillery Tactics
TB-0202	TCAD	90	Artillery Tactics
WC-0211	WD	30	Firing Battery
WC-0212	WD	50	Firing Battery
WC-0214	WD	40	Firing Battery
WM-0213	WD	50	Maintenance Management
WM-0215	WD	30	Maintenance Management
WM-0216	WD	45	Maintenance Mangement
WM-0217	WD	25	Maintenance Management

Once a spanning tree has been developed, the components are hierarchically arranged by the magnitude of the correlations at the connection points to develop an hierarchical tree. A spanning tree and resultant hierarchical tree for eleven selected comporent grades in FAOBC were structured from the FAOBC Sample I data. Those items are graphically portrayed in Figure 1. Since FAOBC Final Grade is a weighted average of 19 component grades, a spanning tree including this score would necessarily be biased in the direction of the component grades clustering close to the final grade. Similarly, if individual shoot scores or data which directly influenced those scores were included, they would be expected to be connected to the GØ-0211 Observed Fire Score.

Inspection of the spanning tree in Figure 1 shows that six of the eleven grades evaluated cluster around the GD-0202 score which, like GD-0203 and GD-0204, is a written gunnery department exam emphasizing fire direction procedures. The four other items connected to the GD-0203 score are two tactics exams, TB-0201 and TB-0202; a communications exam, CC-0201; and a test on the FADAC computer, GD-0205. One should note that the map reading and navigation exam AA-0201, a performance based exam, and the observed fire practical grade, GØ-0211 were not directly connected to the core of the spanning tree.



SPANNING TREE

HIERARCHICAL TREE

Figure 1 SPANNING TREE AND HIERARCHICAL TREE ILLUSTRATING RELATIONSHIPS AMONG CANDIDATE CRITERION MEASURES.

One may also note that in the hierarchical tree that the GØ-0211 score had the lowest connecting correlation of the components of the tree. This is not to imply that GØ-0211 was not related to the other scores. Indeed, one can readily educe from the spanning tree that a relationship between this score and the others existed, but GØ-0211 was not as closely associated with the other items as those items were with each other.

[2] Though models were constructed for other dependent variables as part of the total research project, only models of GØ-0211 performed are reported here.

When multiple correlations of eleven selected grades with FABOC Final Grade were computed, some notable findings were revealed. Since FAOBC Final Grade is a weighted mean of the component grades, one would expect the grades with the strongest weighting factor to have higher correlations with FAOBC Final Grade. Interestingly, the GØ-0211 score had the strongest weighting factor for FAOBC Final Grade but also the second weakest correlation. Because of its heavy weighting in the computation of final grade, one could anticipate a priori GØ-0211 to have one of the highest correlations with FAOBC Final Grade.

Inspection of the spanning tree, hierarchical tree, and correlations with FAOBC Final Grade suggested, for criterion selection, that the GØ-0211 score was likely to reflect a set of skills and abilities which was less redundant with those skills and abilities reflected in the other test scores or the final grade which combined all of them. The practical importance of this finding is that, from a quantitative point of view, the GØ-0211 score emerged as a reasonably strong criterion even when FAOBC Final Grade is to be used as a criterion.

The observed fire grade, GØ-0211, necessarily included some instructor bias. Despite this, the GØ-0211 grade remained the best available FO performance criterion because it was an estimate of FAOBC student firing skills which was less susceptible to extraneous factors than other measures such as raw target location error on firing exercises. The GØ-0211 score necessarily included, to some extent, an element of subjectivity, but this variance was, on the whole, tolerable since there was no suggestion of systematic bias noted in observation of the firing exercises or from student comments on training evaluation questionnaires completed by three FAOBC classes.

Performance Modeling. Following the criterion selection process, several interim analyses were completed prior to the development of the predictive models. These interim analyses were conducted on FAOBC Sample I data. First, an item analysis was completed in which individual items from the FOPPQ were correlated with the GØ-0211 score. Second, factor analysis was performed. In order to hold correlation matrices to manageable sizes and since this was an interim step, three separate factor analyses were performed. A varimax rotation was used. One was completed on Section A, another on scalable items from Sections B andC, and the third on Sections D and E. Items from Sections B and C which were not included in the factor analysis were individually compared according to various scoring schemes to determine which would work best for that item. For example, item C2 which pertained to first branch choice was analyzed with each branchchoice separately; Artillery versus all other branch choices; and Artillery, other noncombat, and a third category consisting of Infantry, Armor, Combat Engineer, Finance. and Adjutant General. For descriptive purposes, individual responses were of value; but for predictive purposes, the categorization scheme, Artillery versus all other branch choices, worked well and thus was used.

The factor analyses were performed to identify redundancies among test items and as an aid in reducing the total set of possible predictors to a more easily manipulated and interpreted set. Just as correlation coefficients for individual items with the GØ-0211 score had been determined during the item analysis activity, correlations between factors and the GØ-0211 criterion measure as well as other FAOBC grades were computed.

In the interest of retaining simplicity, questionnaire items which loaded heavily on certain factors were identified and simple counts of those items were entered into the regression. By using counts instead of factor weights, greater ease in scoring the test was achieved. Instead of requiring the use of a powerful computer, this allows the use of simple scoring keys. If a computer is available, the scoring can still be done by machine. Differences in the predictability of the model with counts versus factor weights were expected to be slight, whereas, gains in simplicity were expected to be great. An example of the simpler technique is evident in the "Sports" score which was used. Six items from Section A of the FOPPQ made up this score. They were A1-8, skeet or trap-shooting; A1-20, hiking; A1-23, golf; A1-25, baseball; A7-4 participating in sports; and A7-5, observing sports. If an individual answered in the affirmative on all of these items, he scored a six on this simple scale.

Some items were analyzed using dummy variables, categorical variables created to specify classes when a continuum is nether available nor appropriate. Dummy variables are frequently used to sort effects of ordinal factors when higher scaling techniques are not appropriate or to sort the effects of a nominal variable such as source of commission. If three categories were chosen, ROTC, USMA, and other, and if an individual were an ROTC officer, he would receive a value of 1 in the ROTC vector and a value of zero for each of the other categories. Thus, in a regression model the difference would be picked up but no ordinal scaling assumptions (which were not appropriate here) would be required. An example of the use of dummy variables in the present modeling effort was the Boy Scouts question which was A3 on Developmental Form A. By using dummy (binary) variables, the impact of each rank in Boy Scouts could be sorted separately without an artificial penalty or benefit for not being in scouts. Such an artificial penalty would have emerged if not being a scout had been assigned a scale value lower than tenderfoot on a dimension of Boy Scout rank.

GØ-0211 was regressed on many combinations of potential predictors prior to selecting the set of predictors which comprise the models reported here. In the model building activity, some predictors effectively displaced others which contained redundant information. It was apparent from even a cursory examination of the data that a mathematical aptitude or ability factor would be an important predictor. This was reflected in the analysis of items such as A4 which asked which mathematics courses had been completed. Having completed calculus correlated with the GØ-0211 score, r=.18. The strong effect of mathematics was also reflected in the Lorge Thorndike intelligence test scores and the quantitative horizon of the Sequential Test of Educational Progress (STEP) scores. When all of the mathematics predictors were in the model, the variance due to this factor was distributed among them. When one or more of those factors was removed, however, the total variance previously accounted for by all of them was explained by the math predictors remaining in the model. After trying several combinations, it became clear that STEP reflected this aptitude as a proxy for other math predictors. It stood in and reflected most of the effects of all of the math predictors. Thus, the STEP score was included in the models reported below.

Models were built and initially validated using data for students in the FAOBC Sample I class and cross validated on data collected from students in FAOBC Sample II and subsequently on data collected from students in FAOBC Sample III. The analysis of the data from FAOBC Sample III also included a comparison of long and short forms of the FOPPQ which is discussed elsewhere. Double cross validation procedures which consisted of separately building the best model for each of two samples and cross validating each model on the other sample were used. Thus a model constructed on FAOBC Sample II data was validated on data collected from FAOBC Sample I.

When comparing the adequacy of fit of two or more models, one could consider as best either the one with maximum R^2 or the one with the minimum standard error. Generally, the two measures will select the same model, but where this would not be true, more importance was assigned to the standard error. This is because the standard error, unlike R^2, takes account of the degrees of freedom of the model error and thereby avoids the pitfall of inflating the estimate of variance attributable to the predictor variables. Thus, the preferred model will be the most robust in cross validation, with respect not only to standard error, but R^2 as well. In other words, the present results are more likely to be repeatable with new data.

Scores below 69 were set to 69. This was done primarily because the discontinuties in the data are exaggerated for the failing grades (below 70) in contrast to the passing grades, suggesting that differences between the failing students are similarily exaggerated (Draper and Smith, 1966). The transformation of the scores also reduces the standard error of the predictive models. This value, 69, was chosen not only because it was one point below passing scores, but also because some exploratory regression analysis showed that predicted scores were rarely below that value.

The first predictive model constructed for the Observed Fire score was

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots\beta_{29}X_{29}+\varepsilon \qquad (1)$$

where Y is the GØ-0211 grade, and the predictors, in order of acceptance into the model, are as shown in Table 2. This model yielded an R^2 of approximately .48. Items from the FOPPQ included in the model were of three types. First was the six point sports scale described earlier, second were items treated as dummy variables, and third were items from Sections D and E of the questionnaire. Items in Sections D and E were scaled on a five point dimension ranging from Strongly Agree (1) to Strongly Disagree (5). A scoring example is provided with a later regression model.

Since Section D and E items are rated from Strongly Agree (1) to Strongly Disagree (5), interpreting Section D and E items in the model at a descriptive level required great care because of the sign associated with the $\hat{\beta}$ value and possible reverse wording in the questionnaire item. Inspection of item D3 serves to illustrate this point. The statement was, "Being an FO is a rewarding job." The sign on the $\hat{\beta}$ in Table 2 is negative which means that the more strongly the student disagreed with this statement, the lower the estimated GØ-0211 score would be. On the average, then, one who strongly disagreed with the D3 statement would be expected to score over six points lower on the GØ-0211 grade than an individual who indicated strongly agree. The summary result for D3 was that those individuals who agreed that being an FO was a rewarding job tended to score higher on the GØ-0211 observed fire grade.

Discussions with intructors at the Field Artillery School suggested that a substantial deficit in map reading and terrain association skills was typical of the FAOBC population. This seemed to require predictor variables which reflected map reading performance. Students are assumed to possess the requisite map reading skills prior to FAOBC training, but it was clear that wide differences existed along this dimension. Ideally, from a personnel selection point of view, one would want a pre-FAOBC measure of map reading ability which would clearly reflect the ability to apply map reading principles to real terrain, i.e., a job sample test.

TABLE 2

SUMMARY OF THE MULTIPLE REGRESSION OF CO-0211
ON A SET OF PREDICTOR VARIABLES REPRESENTING
15 VARIABLE CATEGORIES: FADBC SAMPLE 1

	VARIABLE DESCRIPTION		$\hat{B}$	F AT ENTRY	F	INCREASE IN R^2	MULTIPLE R^2
	STEP	x_1	0.1292	19.03	9.72	.0649	.0649
	SPORTS	x_2	0.9457	20.54	5.51	.0739	.1388
B4	URBAN	x_3	-4.8467				
	SUBURBAN	x_4	-1.2744	6.37	4.55	.0459	.1847
	(RURAL)		0.0				
A3	NO RESPONSE	x_5	-4.9064				
	TENDERFOOT OR SECOND CLASS	x_6	-4.4573				
	FIRST CLASS	x_7	-5.9934	3.99	3.92	.0718	.2565
	STAR OR LIFE	x_8	-3.0124				
	EAGLE	x_9	-0.4164				
	(NO BOY SCOUTS)		0.0				
	D3	x_{10}	-1.6167	7.49	5.54	.0269	.2834
	E2	x_{11}	1.2102	6.81	4.35	.0245	.3079
B25	NO RESPONSE	x_{12}	-12.3447				
	NO SINGLE AREA	x_{13}	-5.2254				
	MATH - SCIENCE - ENGR	x_{14}	-0.7345				
	BIOLOGY - PHYSIOLOGY	x_{15}	3.0504				
	ENGLISH - JOURNALISM	x_{16}	1.7473	2.34	2.29	.0758	.3837
	BUSINESS	x_{17}	0.4151				
	FOREIGN LANGUAGE	x_{19}	-2.6021				
	HISTORY - POLITICAL SCIENCE	x_{19}	-2.9571				
	PSYCHOLOGY - EDUCATION	x_{20}	-3.5917				
	(OTHER)		0.0				
C8	NO RESPONSE	x_{21}	2.3568				
	OTHER THAN REFERENCE	x_{22}	-2.4169	3.54	3.67	.0254	.4091
	(INADEQUATE PERFORMANCE BY FO)		0.0				
	D5	x_{23}	-1.5584	4.57	7.18	.0165	.4256
	D11	x_{24}	0.8645	2.63	3.11	.0095	.4351
	D16	x_{25}	-1.0299	3.44	2.70	.0124	.4475
	D10	x_{26}	-0.8399	2.14	2.77	.0077	.4552
	E5	x_{27}	1.1756	2.54	2.60	.0091	.4643
	E10	x_{28}	0.7929	2.07	2.10	.0075	.4718
C2	ARTILLERY	x_{29}	1.3523	1.60	1.60	.0057	.4775
	(OTHER)		0.0				
	INTERCEPT		49.9407				

TABLE 3

SUMMARY OF THE MULTIPLE REGRESSION OF CO-0211
ON A SET OF PREDICTOR VARIABLES REPRESENTING
16 VARIABLE CATEGORIES: FADBC SAMPLE 1

	VARIABLE DESCRIPTION		$\hat{B}$	F AT ENTRY	F	INCREASE IN R^2	MULTIPLE R^2
	STEP	x_1	0.0767	19.34	3.20	.0649	.0649
	SPORTS	x_2	0.7232	22.04	4.29	.0739	.1399
B4	URBAN	x_3	-4.6301				
	SUBURBAN	x_4	-1.4717	6.94	4.32	.0459	.1947
	(RURAL)		0.0				
A3	NO RESPONSE	x_5	-1.6361				
	TENDERFOOT OR SECOND CLASS	x_6	-3.7240				
	FIRST CLASS	x_7	-5.2319	4.29	2.99	.0719	.2565
	STAR OR LIFE	x_8	-3.9966				
	EAGLE	x_9	-0.8983				
	(NO BOY SCOUTS)		0.0				
	D3	x_{10}	-1.2530	9.03	3.49	.0269	.2934
	E2	x_{11}	1.2184	7.31	4.73	.0245	.3079
B25	NO RESPONSE	x_{12}	-4.6993				
	NO SINGLE AREA	x_{13}	-4.5164				
	MATH - SCIENCE - ENGR	x_{14}	-0.2904				
	BIOLOGY - PHYSIOLOGY	x_{15}	3.1316				
	ENGLISH - JOURNALISM	x_{16}	2.3260	2.51	1.73	.0758	.3837
	BUSINESS	x_{17}	1.2490				
	FOREIGN LANGUAGE	x_{19}	-2.1703				
	HISTORY - POLITICAL SCIENCE	x_{19}	-2.2570				
	PSYCHOLOGY - EDUCATION	x_{20}	-2.1317				
	(OTHER)		0.0				
C8	NO RESPONSE	x_{21}	1.6066				
	OTHER THAN REFERENCE	x_{22}	-2.2642	3.79	3.12	.0254	.4091
	(INADEQUATE PERFORMANCE BY FO)		0.0				
	D5	x_{23}	-1.5772	4.90	7.99	.0165	.4256
	D11	x_{24}	0.9541	2.92	3.26	.0095	.4351
	D16	x_{25}	-1.2193	3.69	4.03	.0124	.4475
	D10	x_{26}	-0.7533	2.30	2.39	.0077	.4552
	E5	x_{27}	0.8394	2.72	1.39	.0091	.4643
	E10	x_{29}	0.7764	2.22	2.16	.0075	.4719
C2	ARTILLERY	x_{29}	1.1713	1.72	1.29	.0057	.4775
	(OTHER)		0.0				
	AA-0201	x_{30}	0.1772	11.56	11.56	.0399	.5163
	INTERCEPT		50.0591				

To develop and administer such a test was beyond the scope of this effort, but a performance-based measure (potential predictor) was selected. This measure was typically recorded on the fourth day of FAOBC training. The utility of this measure as a predictor should have substantial correspondence with the utility of a similar job sample test which could easily and inexpensively be given prior to FAOBC. The measure selected was the score on the map reading practical examination, AA-0201. That test consists of self and target location measures and requires the application of terrain analysis skills.

An independent and very strong argument can be made for using a job sample test such as the AA-0201 measure early in training as a predictor of later performance (training or beyond). This approach has shown very good results. Long and Varney (1975) reported the successful application of the job sample approach to selection in their discussion of a pilot selection program that was developed for the Air Force Human Resources Laboratory at Lackland Air Force Base. A five-hour job sample of flying tasks was administered to 178 candidate flight students that were tested with the automated pilot aptitude measurement system (APAMS). APAMS consisted of two General Aviation Trainers, a minicomputer, and several audio/visual devices for presenting instruction and feedback. Performance measures collected with the APAMS were then correlated with later performance during undergraduate pilot training. Results indicated that performance in all phases of training could be predicted from performance on the "learning sample". Analyses indicated that use of APAMS as a selection tool could reduce attrition rates during UPT from 35 percent to less than 10 percent.

Recently, Marco, Bull, and Vidmar (1978) developed an approach to helicopter pilot selection for the Army called the Proficiency-based Aviator Selection System (PASS) which utilized the job sample technique demonstrated in the APAMS program. An evaluation of the predictive validity of PASS is currently underway and the preliminary results look very promising. Both studies suggest that the use of a simulator coupled with a job sample measurement system is an effective method of selecting individuals with the requisite abilities to learn the task. Additionally, for FAOBC, it may also be an effective way of identifying individuals who may experience difficulty in training in order to provide the additional instruction required for them to successfully complete the program.

<u>Map Reading Predictor</u>. The AA-0201 score was added as a predictor to the predictive model of observed fire performance. The regression of GØ-0211 on the predictors in Equation 1 plus the AA-0201 score yielded the model

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_{30}X_{30}+\ \varepsilon \qquad (2)$$

which is summarized in Table 3. The addition of AA-0201 was not only statistically significant ($p<.01$), but also raised the multiple R^2 to approximately .52. The implications of this performance-based map reading score are important especially in light of other findings regarding the map reading ability of FAOBC students.

As an interim step to finding a sound but not combersome predictive model, the set of predictors was reduced to those three items which had very strong impact in the larger models. The reduced model constructed using the FAOBC Sample I data (3) is summarized in Table 4. The values of R^2 for these two models are .18 and .28 respectively. The same three predictors along with the AA-0201 grade yielded the model

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\ \varepsilon \qquad (3)$$

which is summarized in Table 5. As with the larger models, when the AA-0201 score was added to the model the predictive validity rose sharply.

Clearly model building is an art guided by the results of preliminary statistical analyses but not forced by the. It should be noted that a large number of models could be constructed using the present data, and those models which have been created reflect educated guesses as well as statistical analyses. The value of a model is only established when that model is cross validated on a second independent sample. All models constructed on FAOBC Sample I data were cross validated on FAOBC Sample II and the three and four element models were subjected to double cross validation procedures; the FAOBC Sample I models were validated on FAOBC Sample II data and the FAOBC Sample II models were validated on FAOBC Sample I. Table 6 presents the validity coefficients for both model building and cross validation samples. As would be expected, in the model building stage, those equations with more predictors yielded larger validity coefficients than equations with more predictors yielded larger validity coefficients than equations with fewer elements. When the smaller set is a subset of the larger set, this will necessarily follow. There are no such constraints on the cross validation process since the models are being applied to an independent sample. Normally one expects the validity coefficient R, to be smaller in cross validation than in model building. Generally, less shrinkage is indicative of a more robust model.

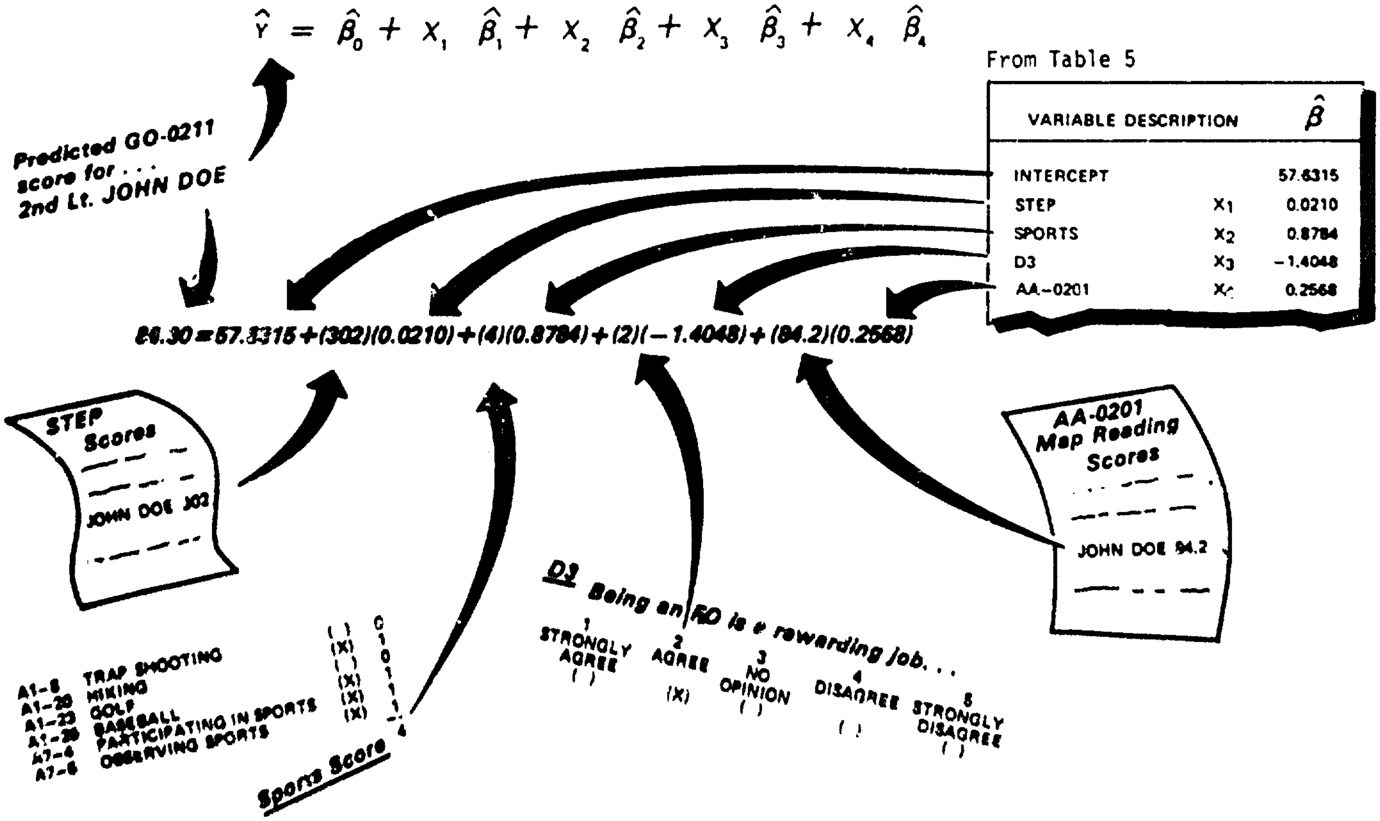
A simple example of how one would compute a predicted score for a particular student is provided for the model described in Table 5. Each obtained score is multiplied times the corresponding β value and those products and the intercept are summed to give a predicted value. Figure 2 illustrates this process and further clarifies handling of component elements. The compoents shown are the sports score and item D3, a five point scale item. The basic procedure illustrated in Figure 2 can be followed in computing a predicted score according to any one of the models presented in this section.

TABLE 4
SUMMARY OF THE MULTIPLE REGRESSION OF GØ-0211
ON THREE PREDICTORS: FAOBC SAMPLE I

VARIABLE DESCRIPTION			F AT ENTRY	F	INCREASE IN R^2	MULTIPLE R^2
STEP	X_1	.1233	13.49	9.45	.0649	.0649
SPORTS	X_2	1.2001	15.57	10.08	.0740	.1389
D3	X_3	-1.8190	7.92	7.92	.0381	.1770

TABLE 5
SUMMARY OF THE MULTIPLE REGRESSION OF GØ-0211
ON FOUR PREDICTOR VARIABLES: FAOBC SAMPLE I

VARIABLE DESCRIPTION			F AT ENTRY	F	INCREASE IN R^2	MULTIPLE R^2
STEP	X_1	0.0210	15.65	.25	.0649	.0649
SPORTS	X_2	0.8784	17.83	6.08	.0740	.1389
D3	X_3	-1.4048	9.18	5.39	.0381	.1770
AA-0201	X_4	0.2568	28.30	28.30	.1175	.2945
INTERCEPT		57.6315				



$$\hat{Y} = \hat{\beta}_0 + X_1\ \hat{\beta}_1 + X_2\ \hat{\beta}_2 + X_3\ \hat{\beta}_3 + X_4\ \hat{\beta}_4$$

From Table 5

VARIABLE DESCRIPTION		$\hat{\beta}$
INTERCEPT		57.6315
STEP	X_1	0.0210
SPORTS	X_2	0.8784
D3	X_3	−1.4048
AA−0201	X_4	0.2568

$$86.30 = 57.6315 + (302)(0.0210) + (4)(0.8784) + (2)(-1.4048) + (84.2)(0.2568)$$

Figure 2 Example of the computation of a predicted score using the four element model for GO-0211 performance.

TABLE 6

Summary of Validity Coefficients (R) for Three Samples*
With GØ-0211**

No. of Predictors in Equation	Model Building		Cross Validation		
	FAOBC Sample I	FAOBC Sample II	FAOBC Sample I Model on FAOBC Sample II Data	FAOBC Sample II Model on FAOBC Sample I Data	FAOBC Sample I Model on FAOBC Sample III Data
16	.691 (.478)	.583 (.340)	.292 (.086)	-	.398 (.158)
15	.719 (.516)	.598 (.357)	.318 (.101)	-	.399 (.159)
4	.543 (.294)	.574 (.299)	.464 (.215)	.468 (.219)	.352 (.124)
3	.177 (.031)	.278 (.077)	.258 (.066)	.163 (.026)	-

*Cell entries are R values. The values in parenthesis are R^2.
**For further clarification see "Selection and Training of Field Artillery Forward Observers: Methodologies for Improving Target Acquisition Skills."

Half of the students in the FAOBC Sample III cross validation sample were administered the FOPPQ Developmental Form A and the other half were administered a shortened version, Developmental Form B. The shortened version was developed to reduce test administration and processing time if the questionnaire were to be widely used, and to ensure that the predictive value of the items was not a peculiarity due to item placement, i.e., context factors. Developmental Form B was created by eliminating some Developmental Form A items which proved to be redundant, items which did not obtain a distribution of scores for the FAOBC population, and items which did not show promise as predictors of FO performance.

Responses obtained from FAOBC Sample III for the two forms of the FOPPQ were compared and Chi Square values were computed when comparing items from the two forms, which were included in any of the models. Item C2 was the only item found to yield significantly different results $X^2(1)=6.6(p<.01)$ for the two forms. Since item C2 required a specification of branch choice, a clear and straightforward question, no practical significance was attached to that difference. For the cross validation on FAOBC Sample III data, a single grouping of students completing Developmental Form A or Developmental Form B was used.

Earlier, the importance of AA-0201, the Map Reading grade, was mentioned. This grade, which was obtained on the fourth day of normal instruction, appeared to be a good predictor of success in FAOBC. The effect was believed to be multifaceted, reflecting motivational components (if one scores low early in the course it might be more difficult to be motivated for later segments), skill components, and, undoubtedly, other factors. Since this test is given so early in FAOBC it is doubtful that it reflects what has bee learned in FAOBC as much as it reflects pre-FAOBC training. This is particularly important in light of the fact that the present Course of Instruction for FAOBC does not include any map reading/land navigation instruction except for a seven hour review of the basics which is conducted by the Counterfire Department during the first three days of FAOBC. One of the recommendations of the USTEA-I (USAFAS, 1977) study was that training of map reading skills in FAOBC be improved. The present findings were consistent with instructor comments that a portion of the lieutenants entereing FAOBC do not have basic map reading, navigation, or terrain association skills. The strength of AA-0201 as a predictor probably comes from its ability to detect this difference early in FAOBC.

The FOPPQ could serve effectively as a personnel selection device, but it also has other applications which are not dependent upon manpower flow issues. Both individualized and group adaptive training approaches might be applied. The FOPPQ can, during FAOBC, serve as an indicator of those requiring additional or modified training in several content areas including map reading. Similarly it could be useful as a counseling aid both at the U.S. Army Field Artillery School and at precommission training sites, e.g., in ROTC,

References

Draper, N. R., & Smith, H. *Applied regression analysis.* New York: Wiley, 1966.

Long, G. E., & Varney, N. C. *Automated pilot measurement system* (AFHRL-TR-75-58). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, September 1975. (NTIS No. AD-A018 151).

Marco, R. A., Bull, R. F., & Vidmar, R. L. *Rotary wing proficiency - based aviator selection system (PASS)* (Final Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, March 1978.

Nelson, A. E., Marco, R. A., & Banks, A. L. Character profile of the West Virginia Coal Mines: Analysis of demographic, personal, attitudinal, and academic achievement data. In *A training analysis approach to the education of coal miners* (Appendix C , MDC E1566). St. Louis, MO: McDonnell Douglas Corporation, August 1976.

U.S. Army Field Artillery School. *Weapon system training analysis - The forward observer, phase 1A, baseline* (Vol. 1, ACN 32750). Fort Sill, OK: Author, May 1977.

Youngling, E. W., Levine, S. H., Mocharnuk, J. B., & Weston, L. M. *Feasibility study to predict combat effectiveness for selected military roles: Fighter pilot effectiveness* (MDC E1634). St. Louis, MO: McDonnell Douglas Corporation, April 1977.

ASSIGNMENT OF ARMY AVIATOR TRAINEES TO UNDERGRADUATE AEROSCOUT MISSION TRAINING

John A. Dohme, Ph.D.
Army Research Institute Field Unit
Fort Rucker, Alabama

INTRODUCTION

Army aviators are first trained in helicopter flight in the Initial Entry Rotary Wing (IERW) Course at the US Army Aviation Center (USAAVNC) at Fort Rucker, Alabama. In June 1977, the format of the IERW Course was changed. Previously, trainees all received their wings in the Utility mission flying the UH-1 aircraft. The new format added an additional training track in which 25% of IERW graduates earned their wings in the Aeroscout mission utilizing the OH-58 aircraft. The US Army Research Institute for the Behavioral and Social Sciences (ARI) Field Unit at Fort Rucker was tasked to develop and validate a method of identifying trainees for Aeroscout assignment.

Additionally, ARI has been tasked to develop means of identifying trainees for assignment to a multi-track IERW Course in which they would earn their wings in one of four helicopter missions: Aeroscout, Attack, Cargo or Utility. This effort is underway and ARI is awaiting the USAAVNC decision on whether the multi-track format will be adopted. The Aeroscout selection methodology reported in this paper is essentially an interim solution to the proposed multi-track objective and an opportunity to develop the necessary technology for the multi-track assignment problem.

METHOD

Algorithm Development and Application

Subject Matter Experts (SMEs) representing three Army Aviation missions in the field defined the criteria for Aeroscout mission proficiency. Experienced aviators in the Aeroscout, Attack and Utility missions were interviewed to identify selection criteria that would discriminate successful Aeroscout aviators. Information from these interviews was used to create a questionnaire containing 26 characteristics which were expected to differentiate among the three missions. The questionnaires were administered to 120 experienced Aeroscout, Attack and Utility aviators serving with the 6th Cavalry Brigade (Air Combat) at Fort Hood, Texas. Each aviator ranked the relative importance of each of the 26 characteristics for each of the three missions. The selection criteria for Aeroscout aviators were determined from the rankings averaged across the 120 SME aviators.

The selection criteria for Aeroscout mission proficiency became predictors in the Aeroscout selection algorithm. Table 1 presents the nine field derived predictor variables, their shorthand identifiers, the source of the data for that variable and the weights used to input the relative importance of that variable to the algorithm. The weights were determined from the average importance rankings. Variables 10 and 11 (RANK and PGRADE) were not derived from the questionnaire data but were included as general estimates of the trainee's performance in the IERW Course.

The algorithm is a mathematical combination of the eleven predictor variables collapsed across raters and weighted by the field-derived coefficients. It is normed to an approximate mean value of 85 in order to be directly comparable to IERW training grades. The algorithm value (also termed the Aeroscout Final Composite Score) is a major component in the decision model used by USAAVNC to assign trainees to the Aeroscout Track. The decision model first screens trainees on these criteria, which are Department of the Army (DA) regulations and policy:

1. Component of the Army - Must be Regular Army
2. Sex - Must be male
3. Branch of Service - Officers must be Combat Arms
4. IERW Course Progress - Must not be a setback

TABLE 1 - VARIABLES WHICH SERVE AS PREDICTORS IN THE AEROSCOUT ALGORITHM

NO.	VARIABLE NAME	VARIABLE IDENTIFIER	SOURCE	WEIGHT
1	Map Reading Skills	MAP	IP Evaluation	.475
2	Sense of Direction	SDIR	IP Evaluation	.453
3	Flying Ability	FLYING	IP Evaluation	.428
4	Leadership Abilities	LEAD	IP and TAC Officer Evaluation	.366
5	Performance Under Stress	STRESS	IP and TAC Officer Evaluation	.463
6	Teamwork Ability	TEAM	IP and TAC Officer Evaluation	.444
7	Ability to Divide Attention	DIVATN	IP and TAC Officer Evaluation	.434
8	Verbal Expression	VERBAL	IP and TAC Officer Evaluation	.394
9	Aggressiveness	AGG	IP and TAC Officer Evaluation	.391
10	Standing in a Typical 25 Trainee Group	RANK	IP and TAC Officer Evaluation	
11	Average Grade at 14th Training Week	PGRADE	IP and TAC Officer Evaluation	

IERW Course trainees are rated on the predictor variables by either two or three individuals depending on whether the ratee is a Commissioned Officer or a Warrant Officer Candidate (WOC). Each officer is rated by a Primary Phase Instructor Pilot (IP) at about the tenth training week and by a Contact Phase IP at about the fourteenth training week. The WOCs are also rated by the Training, Advising and Counseling (TAC) officer by the fourteenth training week.

Then eligible trainees are assigned to the Aeroscout Track based conjointly on algorithm scores and trainee mission preference.[1] Assignments are made subjectively by Student Personnel Operations Section (SPOS) personnel at USAAVNC considering first the trainee's preference and then the algorithm score.

Algorithm Validation

The predictive validity of the Aeroscout selection process has been estimated using both institutional and operational criteria. IERW Course grades obtained from trainee flight folders at USAAVNC served as institutional criteria. Aeroscout trainees take 19 academic exams, 7 flight exams (checkrides) and are evaluated on their flight/mission abilities by 7 IPs. Of special interest in this validation effort are the Aeroscout Tactics grades which include 3 academic exams, 2 checkrides and 2 IP evaluation grades (on transition to the OH-58 aircraft and on Aeroscout mission tactics).

The operational criterion was difficult to develop. Estimating the mission proficiency of Aeroscout Track graduates serving in Army field assignments requires a sampling of their mission flight skills. This, in turn, requires expensive aircraft and IP time to perform the flight/mission evaluation. ARI accomplished the mission proficiency assessment objective by cooperating with the Flight Standardization Division of the Directorate of Evaluation and Standardization (DES) at USAAVNC. It was considered to be an efficient use of Department of the Army (DA) funds to instruct the Standardization Instructor Pilots (SIPs) to perform the Aeroscout proficiency assessments as an adjunct to their usual mission of giving field standardization checkrides. When the SIPs went to the field, they searched the rosters for recent Aeroscout graduates and attempted to evaluate as many of them as possible.

A mission proficiency instrument was developed in cooperation with the Standardization Instructor Pilots (SIPs) at DES. The instrument uses a decision tree model adapted from the Cooper-Harper scale which was devised to rate aircraft handling qualities (Cooper, 1957; Harper and Cooper, 1966). The Mission Proficiency Scale (MPS), which is presented in Figure 1, yields a single score reflecting the aviator's overall mission proficiency. The logic of the MPS is based on the assumption that proficiency in FAC 2 (aircraft control) tasks is basic to proficiency in FAC 1 (mission tactics) tasks. The rationale behind the MPS is to allow an SIP to evaluate an aviator's mission proficiency without directing the SIP's attention away from the flight, i. e., without requiring in-flight note taking. The MPS form can be completed during the post-flight debriefing thus, it does not compromise aircrew safety.

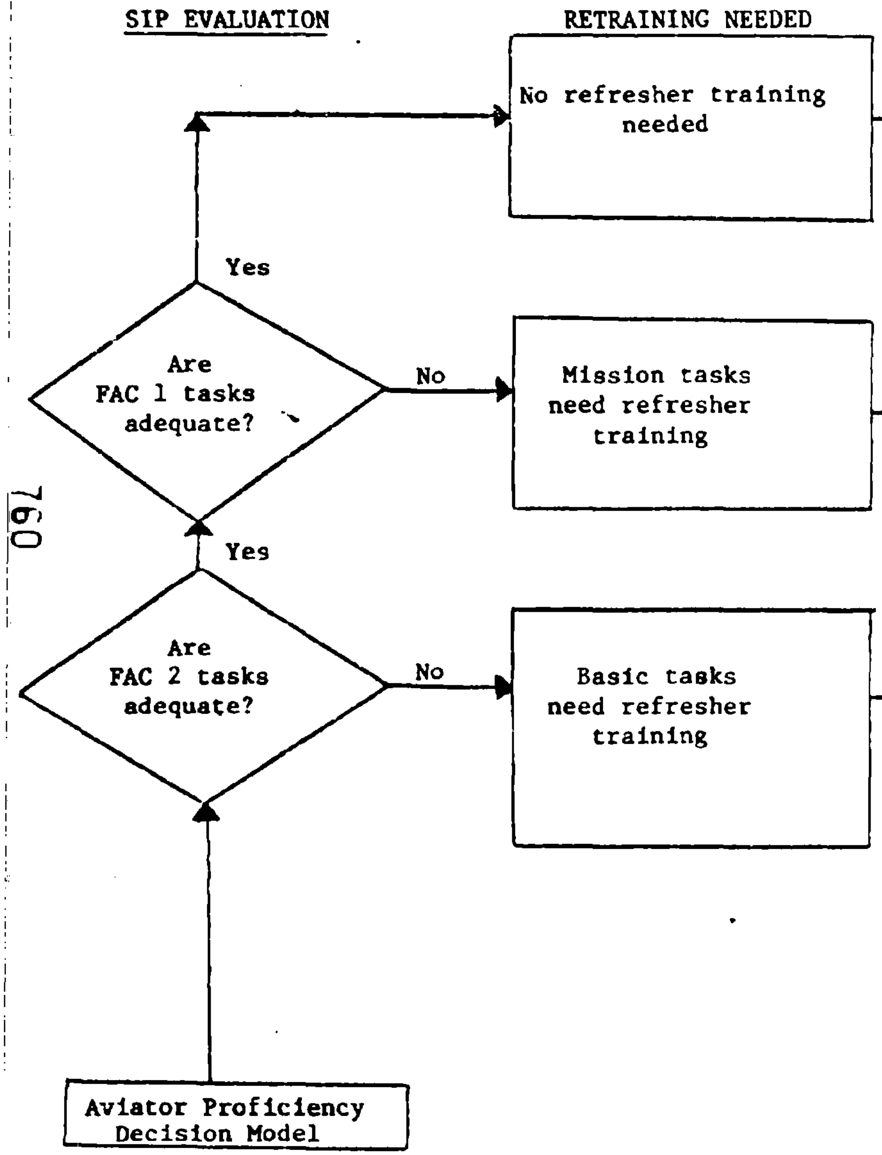
Subjects for the validation study were the first 248 Aeroscout Track graduates. A subset of 15 of those subjects served as the operational validation sample. The DES SIPs used the MPS to evaluate any Aeroscout Track graduates they were able to locate as part of their standardization mission. Over a 5 month period, 15 such aviators were located and tested. Additionally, a sample of 52 Utility

[1] Each trainee responds to an ARI developed instrument which describes the four basic helicopter missions and asks the trainee to rank order them in terms of preference from one to four.

FIGURE 1 - MISSION PROFICIENCY SCALE (MPS)

AVIATOR'S NAME ____________________

AVIATOR'S MISSION ASSIGNMENT ____________________


SIP EVALUATION
RETRAINING NEEDED
Aviator Proficiency Decision Model
Are FAC 2 tasks adequate?
No
Basic tasks need refresher training
Yes
Are FAC 1 tasks adequate?
No
Mission tasks need refresher training
Yes
No refresher training needed

AVIATOR PROFICIENCY	AVIATOR RATING (Check one)
Aviator highly proficient at this mission	1
Aviator moderately proficient at this mission	2
Aviator marginally proficient at this mission	3
Basic tasks adequate but 1 or 2 mission procedures out of tolerance	4
Basic tasks adequate but several mission procedures out of tolerance	5
Basic tasks adequate but many mission procedures out of tolerance	6
One or two basic procedures out of tolerance	7
Several basic procedures out of tolerance	8
Many basic procedures out of tolerance	9
Nearly all basic procedures out of tolerance	10

Track aviators was assessed as a control group. One Utility Track trainee was randomly selected per class from those trainees who met the Aeroscout screening characteristics.

Of the 248 Aeroscout subjects, 101 are officers and 147 are WOCs. The Utility subjects include 25 officers and 27 WOCs.

RESULTS

The correlations between the algorithm score and IERW Course grades (for the Aeroscout Track, there are 18 academic grades, 7 IP evaluations and 7 checkrides) are all positive and all but 1 are significant. The lone exception is the exam covering UH-1 cockpit procedures. Summarizing these relationships, the algorithm score correlates with the summed IERW grades as follows:

	Pearson r	Significance
Academic exams	r = .42	p = .001
IP evaluations	r = .54	p = .001
Checkrides	r = .46	p = .001
Overall IERW grade	r = .62	p = .001

However, these correlations are inflated to the degree that one component of the algorithm is the trainee's grade at the 14th training week. Thus, there is common variance in the algorithm score and the IERW grade criteria. When that common variance is removed statistically, the partial correlations of algorithm score and IERW grades are considerably diminished because the summed grades are highly correlated with the training grade at the 14th week:

	Partial r	Significance
Academic exams	r = .19	p<.002
IP evaluations	r = .33	p<.001
Checkrides	r = .37	p<.001
Overall IERW grade	r = .27	p<.001

The partial correlations are more appropriate estimates of the relationships between algorithm scores and IERW grades.

Comparisons can be made between the Aeroscout and Utility samples. While these comparisons are not essential to the validation effort, they do provide indirect evidence that the Aeroscout trainees excel in the IERW Course. Reference Table 2, the six combined means for the Aeroscout sample are all significantly higher than the corresponding Utility sample means. Using the t statistic to test those mean differences, the following values obtain: Algorithm score, t = 4.96 (p<.001); IERW overall grade, t = 5.17 (p<.001); academic exam average, t = 2.98 (p<.005); IP evaluation average, t = 5.44 (p<.001); checkride average, t = 5.28 (p<.001) and tactics phase average, t = 6.10 (p<.001). The assumption behind these comparisons is that Aeroscout Track grades are comparable to Utility Track grades. The Aeroscouts are trained in different tactics by different IPs in a different aircraft. However, the assumption is conservative insofar as the higher scoring Aeroscout trainees are transitioned to a new aircraft and trained in a larger number of tactics during that 8 week phase than are the Utility trainees, thus, it could be anticipated that their training is more rigorous and demanding.

TABLE 2 - MEANS AND STANDARD DEVIATION FOR THE AEROSCOUT AND UTILITY SAMPLES

		AEROSCOUT			UTILITY	
VARIABLE NAME	Officer n = 101	WOC n = 147	Combined n = 248	Officer n = 25	WOC n = 27	Combined n = 52
Aeroscout Algorithm	Mean = 82.2 SD = 4.8	Mean = 81.5 SD = 3.5	Mean = 81.8 SD = 4.1	Mean = 80.7 SD = 5.4	Mean = 76.7 SD = 3.1	Mean = 78.6 SD = 4.8
IERW Overall Grade	Mean = 88.8 SD = 2.3	Mean = 88.0 SD = 1.9	Mean = 88.3 SD = 2.1	Mean = 87.4 SD = 2.5	Mean = 85.9 SD = 2.0	Mean = 86.6 SD = 2.4
Academic Exam Average	Mean = 93.6 SD = 3.2	Mean = 90.6 SD = 3.2	Mean = 91.8 SD = 3.5	Mean = 91.9 SD = 3.7	Mean = 88.6 SD = 2.6	Mean = 90.2 SD = 3.6
IP Evaluation Average	Mean = 87.5 SD = 2.4	Mean = 87.7 SD = 1.7	Mean = 87.6 SD = 2.0	Mean = 86.1 SD = 2.4	Mean = 85.6 SD = 2.3	Mean = 85.9 SD = 2.3
Checkride Average	Mean = 86.9 SD = 2.4	Mean = 86.6 SD = 2.2	Mean = 86.9 SD = 2.3	Mean = 85.6 SD = 3.0	Mean = 84.6 SD = 2.5	Mean = 85.0 SD = 2.8
Tactics Average	Mean = 90.5 SD = 2.7	Mean = 89.1 SD = 2.6	Mean = 89.7 SD = 2.7	Mean = 88.6 SD = 4.3	Mean = 86.0 SD = 2.7	Mean = 87.3 SD = 1.9

Comparisons can also be made between the officer and WOC groups within the Aeroscout sample. Reference Table 2, 5 of the 6 means are higher for the officers. The lone exception is IP evaluation grades which are higher in the WOC group. However, only 3 of the mean differences are significant and in those 3 cases, the officers have higher scores: IERW overall grade, $t = 2.95$ ($p<.005$); academic exam average, $t = 6.66$ ($p<.005$); and Aeroscout tactics average, $t = 3.89$ ($p<.005$).

The predictive validity of the algorithm can also be compared across groups. Within the Aeroscout sample, the officers and WOCs can be compared in terms of the predictions to IERW overall grade and to Aeroscout tactics:

Criterion Variable	Officers $n = 101$	WOCs $n = 147$	z test
IERW Overall Grade	$r = .70$ $p<.001$ $Z = .86$	$r = .55$ $p<.001$ $Z = .62$	$z = 1.83$ $p<.07$
Aeroscout Tactics	$r = .49$ $p<.001$ $Z = .53$	$r = .29$ $p<.001$ $Z = .30$	$z = 1.76$ $p<.08$

The correlation coefficients may be transformed to standard scores using Fisher's r to Z transformation (see Guilford and Fruchter, 1973, p. 146). These Z scores may then be compared using the z statistic. Both of these comparisons approach significance but from these samples, it may be concluded that there is no significant difference between officers and WOCs in the algorithm's prediction of IERW and tactics grades.

Since the algorithm and trainee mission preference are both used to assign Aeroscouts, it is informative to compare them as predictors of institutional performance. The mission preference variable can take four values depending on whether the trainee ranked the Aeroscout mission first, second, third or fourth in preference. The ranks were inverted in order that preference would correlate positively with performance. The following presents the correlations of algorithm score and mission preference with IERW Course and Tactics grades:

	Aeroscout Preference	Algorithm Score
IERW Overall Grade	$r = -.04$ $r^2 = .002$ p = NS	$r = .62$ $r^2 = .38$ $p<.001$
Aeroscout Tactics	$r = .13$ $r^2 = .02$ $p<.04$	$r = .39$ $r^2 = .15$ $p<.001$

The correlations of the algorithm with both criteria are significant whereas only one of the mission preference correlations attains significance.

Another way to view the relative contributions of mission preference and algorithm scores to trainee performance is to perform an analysis of variance treating the two factors as independent variables. Mission preference was assigned to two levels: wants Aeroscout, doesn't want Aeroscout. Algorithm scores were assigned 3 levels; highest 1/3, middle 1/3 and lowest 1/3. The dependent variable is Aeroscout tactics grades. The results of the ANOVA are presented in Table 3 along with omega squared estimates of the amount of variance in tactics grades accounted for by each of the independent variables. Both main effects are significant but the two-way interaction is not. That is, higher algorithm scores and preference for the mission are both related to higher tactics grades and their conjoint effect is additive rather than interactive.

With the relatively large Aeroscout sample size, these F ratios may or may not account for a practical proportion of the tactics variance. To estimate that quantity, omega squared estimates were calculated for both main effects. The algorithm accounts for just over 12% of the variance while Aeroscout preference accounts for less than 1½%. Similar estimates are provided by the r^2 values presented: for the algorithm, $r^2 = .15$ while the value for mission preference is $r^2 = .02$. Thus, the mission preference variable is significantly related to Aeroscout tactics grade but it accounts for little variance. The algorithm does account for sufficient variance to have practical importance in accounting for Aeroscout tactics performance.

The results of the operational validity estimate must be interpreted with caution because the sample size is very small (n = 15). However, the correlation of the algorithm with the Mission Proficiency Variable (MPS) did reach statistical significance ($r = .57$, $p<.05$). MPS also correlated significantly with Aeroscout tactics grade ($r = .56$, $p<.05$) suggesting that the two criteria demonstrate common variance. A stepwise multiple regression of the algorithm components on MPS reached significance for the first two steps. R = .71 with the two components verbal expression and primary grade.

The results of the institutional validity estimate are based on a larger sample size (n = 248). The primary estimate of institutional validity is the correlation of the algorithm with Aeroscout tactics grades ($r = .39$, $p<.001$). The contributions of the individual algorithm components to the prediction of institutional performance can be assessed by means of a multiple regression. The overall F is significant with all 11 components in the equation and R = .48.

However, the only variables contributing significantly based on the F ratio to enter that variable are primary grade, verbal expression, and map reading skills. Table 4 presents the results of the stepwise multiple regression including the beta weights for each of the components.

Since the algorithm is basically a classification tool rather than a selection tool, its validity can also be estimated by performing a discriminant analysis treating the Aeroscout and Utility samples as groups.

The ability of the algorithm to predict group membership can be estimated by classifying the 298 aviators in the known sample. The probability of group membership was set proportional to the two group sizes (Aeroscout 83%; Utility 17%). The results of the classification analysis are presented below:

TABLE 3 - ANOVA SUMMARY TABLE FOR AEROSCOUT TACTICS GRADE TREATING MISSION PREFERENCE AND ALGORITHM SCORE AS INDEPENDENT VARIABLES

SOURCE OF VARIANCE	SS	DF	MS	F	P	Omega2
Mission Preference	1407.27	1	1407.27	4.78	.03	.014
Algorithm	10003.96	2	5001.98	16.98	.001	.122
2 Way Interaction Preference by Algorithm	832.46	2	416.29	1.41	NS	-
Residual	65404.24	222	337.96	-	-	-

TABLE 4 - STEPWISE MULTIPLE REGRESSION OF ALGORITHM COMPONENTS ON AEROSCOUT TACTICS GRADE

	Beta	Multiple		Change in	
Algorithm Component	Weight	R	R^2	R^2	Overall F
1. Primary Grade	.297	.37	.13	.135	36.6
2. Verbal Expression	-.185	.40	.16	.023	21.8
3. Map Reading Skills	.197	.43	.19	.029	17.8
4. Flying Ability	.082	.44	.20	.012	14.3
5. Leadership Ability	-.165	.46	.21	.011	12.2
6. Standing in a Typical Group	.201	.47	.22	.013	10.9
7. Teamwork Ability	-.107	.48	.23	.007	9.7
8. Aggressiveness	-.086	.49	.23	.001	8.5
9. Ability to Divide Attention	.097	.48	.23	.001	7.6
10. Sense of Direction	.028	.48	.23	.000	6.8
11. Performance Under Stress	-.037	.48	.23	.000	6.2

	Predicted Group Membership	
	Aeroscout	Utility
Aeroscout Trainees (n = 247)	96.4% n = 238	3.6% n = 9
Utility Trainees (n = 51)	82.4% n = 42	17.6% n = 9

Overall Correct Classification = 82.9%

The data suggest that subsequent classifications using these discriminant functions will produce a number of false positives but few false negatives. This outcome is not surprising in that a sizable number of trainees who have high scores on the algorithm express a preference for the Utility Track and are assigned to it.

Ten of the eleven components met the criterion for inclusion into the stepwise discriminant function analysis that the change in Rao's V be greater than 0.0. However, only performance under stress and primary grade produced significant changes in Rao's V. The canonical correlation between the components and the two groups is significant (R = .38 p <.001) with the first ten components in the equation. The two group centroids in this analysis are Aeroscout = -.19 and Utility = .88. The standardized discriminant function coefficients, stepwise changes in Rao's V and associated significance levels are presented below:

Component	Coefficient	Change in Rao's V	p
1. Performance Under Stress	-.315	21.95	<.001
2. Primary Grade	-.606	18.65	<.001
3. Sense of Direction	-.242	3.09	.08
4. Aggressiveness	-.525	2.05	.15
5. Ability to Divide Attention	.305	.81	.37
6. Standing in a Typical Group	-.134	.23	.64
7. Map Reading Skills	-.073	.28	.60
8. Verbal Expression	.111	.18	.67
9. Flying Ability	-.083	.14	.71
10. Leadership Ability	-.027	.02	.91
11. Teamwork Ability	.011	Not in analysis	

DISCUSSION

The Aeroscout algorithm has been demonstrated to be an effective assignment tool. It evidences moderate degrees of predictive validity with Aeroscout tactics grade (r = .39) and with IERW overall grade (partial r = .27). The higher correlation with Aeroscout tactics indicates that the algorithm predicts better to specific Aeroscout criteria than to general IERW performance. This can be interpreted as support for the specificity of the algorithm in performing its intended function.

The correlation of the algorithm with the MPS (r = .57) suggests that the instrument may be effective as a predictor of operational performance. However, given the small sample size (n = 15), this conclusion must await further data for confirmation. It is anticipated that DES/ARI will continue the MPS data collection and that a sample of sufficient size to perform a stable multiple regression analysis will be available by mid FY 80.

While the predictive validity estimates are not as high as one might hope for, they are sufficiently large to justify continued use of the current Aeroscout assignment method. The correlations presented above are diminished by the restriction in range of the Aeroscout sample on algorithm scores and on tactics grades. It is difficult to estimate the unrestricted correlation coefficients since all Aeroscout trainees have been selected for assignment based on the decision model presented in the Method section of this paper. Thus, the validity estimates are considered to be conservative.

Additionally, the current Aeroscout algorithm is an interim step toward a multi-track IERW Course in which all graduates will be helicopter mission specialists. It represents an intermediate stage in the development of a technology to test and assign IERW trainees to specific mission based on their aptitudes and abilities. It is anticipated that a multi-track IERW Course will save the Army considerable dollars and man-years in producing combat-ready aviators.

It will be recommended that USAAVNC continue to use the current assignment methodology while ARI continues to consider refinements to the algorithm. The component weighting coefficients in the algorithm could be modified to optimize its predictive validity based on several different strategies:

1. Optimize selection for performance in Aeroscout tactics.
2. Optimize selection for performance in the operational environment.
3. Optimize selection to discriminate between known Aeroscout and Utility trainee characteristics.
4. Optimize selection of commissioned officers and WOCs using different algorithms since the groups differ on a number of dimensions.

These alternatives can be evaluated using a cross-validation sample. ARI intends to devise alternative algorithms and to compare their predictive validities using the above strategies. It is anticipated that this cross-validation effort will be completed in late FY 80.

REFERENCES

Cooper, G. E. Understanding and interpreting pilot opinion. Aeronautical Engineering Review, 1957, 3, 47-56.

Guilford, J. D. and Fruchter, B. Fundamental statistics in psychology and education, 5th edition, New York, McGraw-Hill, 1973.

Harper, R. P. Jr., and Cooper, G. E. A revised pilot rating scale for the evaluation of handling qualities. CAL Rep. No. 153, September, 1966.

Hays, William L. Statistics. New York: Holt, Rinehart and Winston, 1963.

PRACTICAL SOLUTIONS TO CRITERION PROBLEMS IN CBM-X TEST VALIDATION RESEARCH

Major R.T. Ellis and Lieutenant D.A. Saudino
Canadian Forces Personnel Applied Research Unit, Toronto, Canada

This paper concerns a method whereby the Selection and Classification Research team at the Canadian Forces Personnel Applied Research Unit dealt with a problem in validation research. Since it represents only a small part of a larger research project, it is necessary to outline some of the history of the overall project in order to place the problem in context.

The program involves the development of a new selection and classification battery for use in the Canadian Forces. With the advent of the unification of our Army, Navy and Air Force in the late sixties, a set of selection tests, drawn from the batteries used by the three former services, was gathered and entitled the "Classification Battery - Men," or CBM. The various pressures generated by unification delayed any real attempt to validate the CBM until after 1970. Rampton, Skinner and Keates (1972) set about the task, and in the course of their work, identified a number of deficiencies in the old tests. Several of these were dropped from the battery. A new selection model was developed using the remaining tests which optimized the predictive power of the interim battery or CBM-I, as it has come to be called. However, it was clear that a more modern set of tests would be required.

The construction of new aptitude tests is an expensive and time consuming process under any circumstances. For a small force such as Canada's, with a relatively low volume of trainees on which to do large scale test development, it was necessary to decide whether it was within our resources to construct a new battery from the ground up.

Although time and expense were not the only factors, it was eventually decided to undertake a search for an existing battery that could be adapted to meet our needs. Our review revealed the US Army Classification Battery (ACB) as the most likely prospect for Canadian purposes, and permission was secured to begin experimental work with the ACB in 1974. Because Canada is a bilingual country with a bilingual force, the test was translated into French prior to any research being started. Only eight of the ACB sub-tests were used, and in most cases, these were derived by combining a subset of items selected from both forms of the ACB made available to us. The new battery was entitled the Classification Battery Men - Experimental, or "CBM-X".

After initial item analysis and reliability studies, the test was administered to approximately 13,000 recruits. This comprised the complete recruit intake at our two recruit schools, over a 17 month period in 1974/1976 at the English language recruit school, and a 24 month period in 1975/1977 at the French language recruit school.

The criterion selected for validation was performance at the end of initial trades training. This training, in some 67 entry level trades, is conducted at various bases from coast to coast. The progression of training calls for 11 weeks recruit training and up to 33 weeks initial trades training. For unilingual Francophone recruits, up to 24 weeks of additional English language training may intervene to prepare them for trades training in that language.

A criterion data form was devised that was to be filled out on all initial trades trainees at the end of their course. The process of matching CBM-X test scores with criterion data forms was to be done by computer at our research unit after keypunching of the data. The criterion form was designed to gather a rather large number of different types of criterion measures. Its complexity was one of the major contributing factors to problems which the team experienced later.

One of these problems concerned the fact that only three of the types of measures sought on the form were both consistent and available in any numbers within the various trades. These were: Rank-in-Class (stood __ in a class of __ students), Grade assigned, (A,B,C or F) and Disposition (Pass, Fail, Reassign, Recourse, Release). Since the form was designed to be filled out by course instructors at the end of the course, it happened that in the case of trainees who did not complete training, almost invariably no Rank-in-Class was computed, nor was a grade assigned. However, almost all trainees were assigned to one of the Disposition categories. This was likely due to the fact that it would have been relatively simple for instructors to recall at the end of the course what administrative action had been taken on trainees who did not complete training, whereas assigning a grade or rank standing would have been quite difficult. The reasons for this failure to acquire the various kinds of criterion measures intended by the original team are complex, and we have speculated a great deal on what they might be in our attempts to design a mechanism for revalidating the new battery after installation. However, a discussion of these speculations is not within the scope of this paper.

It had been decided early on that our validation analysis would be conducted primarily on a within-trade basis. Given the difficulties engendered by having two different language and culture groups, and the necessity to validate the battery separately for the increasing numbers of women coming into the CF, it was most important to maximize in whatever way possible the number of subjects available in each trade (since our trade X sex X language groups would often be limited in size). The Disposition variable almost always provided us with the largest sample, but initial analysis showed that for use as a nominal variable, there was far too great a proportion of trainees in the "pass" category (normally more than 80%) and too small a proportion in the remaining categories. This rendered it unsuitable as a criterion measure. (It should be noted that our sample did not include those trainees who failed initial recruit training or who may have dropped out during language training ie, 80% or more of those who reached trades training passed). Smaller samples, resulting from failure to record Rank-in-Class and Grade data even on passing trainees, restriction of range problems and a number of inconsistencies and anomalies within the grading system, rendered the Rank-in-Class and Grade variables unusable in many trades.

It soon became apparent that a need existed to combine the information available from one variable on passing trainees (Grade) with the information on unsuccessful trainees on another variable (Disposition). Our strategy was to construct a new variable, which we called the "consolidated criterion", by breaking down the "pass" category of the disposition variable by grade. Since it was considered desirable to be able to conduct multiple regression analysis on this data, the notion of structuring this new variable in such a way as to be amenable to this kind of statistical treatment was very attractive. In reflecting on what the administrative actions represented by the disposition categories might signify in terms of course performance, the idea occurred that a rationale might be developed for placing the categories of the new criterion variable in a meaningful sequence. This, in turn, would allow us to treat it as an ordinal variable. The rationale eventually developed was this:

a. Persons in the "Fail" disposition category were to be assigned a score of 1 on the consolidated criterion. It was assumed that those assigned to this category had likely completed all or most of their training, and had been assessed as unsuitable for further training primarily because of poor performance;

b. Persons in the "Release" disposition category were to be assigned a score of 2. Past experience of the research team indicated that, by and large, this was the most likely action to be taken when a candidate was doing poorly on a course and was prematurely removed from training, either voluntarily or compulsorily. Although our samples undoubtedly contained a number of trainees who were released for reasons not related to aptitude, on the average it was felt that the probability was higher for the reasons behind a given release to be ability-related than would be the case for other disposition categories. Motivational medical and other reasons would presumably underlie other disposition categories with greater frequency.

c. Persons in the "Reassignment" category were to be assigned a score of 3. Generally, an informal policy is known to exist in many CF trades training schools whereby only those trainees who have shown at least a reasonable degree of potential in training are recommended for reassignment. Reassignment is more often used for personnel who have displayed the necessary ability and who want to stay in the service, but who have become disenchanted with the originally assigned trade. This is contrasted with release, which is more likely to be the action taken when the trainers judge the potential for training success to be low, or adjustment to military life to be poor.

d. Persons in the Recourse category were to be assigned a score of 4. Recourses can happen for a number of reasons (medical, compassionate leave, marginal performance, etc.) but the common element in each decision to recourse is the judgement that the trainee can succeed on a second attempt, i.e., he has a reasonable level of ability. In all four of the categories discussed so far, the notion underlying the ordering has been one of probability based on proportion. The key question underlying the rationale is therefore, what proportion of the group defined by a given administrative disposal action is likely to have been assigned to that group because of poor training performance due to low ability? The proportion, it is proposed, varies systematically with each of our four groups. Thus, the probability of assigning a correct ordinal value to an individual in terms of his actual course performance is maximized by using administrative disposal action as the basis for scoring, to the extent that the above assumptions are met.

e. The remaining group of trainees, those who passed, were assigned scores of 5, 6, or 7 according to whether they achieved a C,B, or A grade.

It is not intended to present the results of any validation analysis here, except to state that the results of our analysis of the group of trades which the new variable made available to us were reasonably satisfactory. The criterion data problems identified early in the analysis suggested that we might need to go back to the beginning, and gather a new set of criterion and test data, avoiding the pitfalls encountered in the initial attempt to acquire data. Our results indicated to us that this was not necessary. Because data were available on the old battery as well as the experimental battery, we were able to determine that the new tests were, by and large, performing as well as or better than the old ones, in spite of our data problems. Analysis of the smaller samples available on the remaining entry level trades has been carried out, and an overall selection model devised. Our strategy from this point will be to implement the new battery, and to immediately begin a program of re-validation and "fine-tuning" of the selection model.

In the course of validating the new battery, a number of lessons have been learned, and the process of re-validation will profit from these experiences. It became clear, for instance, that while we know what happened to trainees who were unsuccessful, we are unable to determine why they were unsuccessful. It is virtually certain that there were some high-ability trainees who were performing adequately (or presumably could have were they to have been so motivated) who were released or re-assigned. These are contrasted with those who were released or re-assigned because they simply lacked the requisite ability to perform. For proper validation, we need to be able to identify and separate these groups. Data which would allow this kind of identification is routinely gathered by the CF Training System (CFTS), but the data is in a contaminated and rather crude form for research purposes. We have been successful in assisting the CFTS in setting up a more comprehensive and interpretable set of "disposition action" and "reason for disposition" codes to be used in their recruit and trades training files. Access to this data should allow us to remove from future validation samples those trainees whose failure resulted from reasons not related to the constructs which the tests purport to measure, eg. motivational failures will be excluded from validation samples for aptitude tests, but included in samples used for in the development of interest/personality tests.

Another lesson concerns the problems inherent in attempting to gather complex criterion data on a large scale. Since validation analysis was, and will be, done on a within-trade basis, it is much more important to ensure consistency of criterion data from course to course within a given trade than it is to gather kinds of criterion measures that are similar across trades. We learned that some schools prefer to use Grades as the basis of their assessment scheme while others prefer Rank-in-Class as the most appropriate method of expressing the relative level of training performance. Attempts to impose a non-preferred method of evaluation were the major source of within-trade inconsistency in our earlier data. Thus, our re-validation program has, as one of its objectives, the acquisition of less complex, more consistent data. Modifications have been made to the common CF Course Reporting Form which allow for the recording of several kinds of criterion measures. Schools can thus record whatever type of criterion measure best suits their particular needs. Gaps occurred in some of our CBM-X samples when temporary (summer increment) staff failed to fill out the criterion data form. Since, in effect, the common Course Report will now serve as the criterion data collection form, such gaps should in future be either avoidable, or capable of being filled from archival data.

Perhaps our greatest lesson has concerned the need for proper monitoring of the in-flow of data. For a project of the size just completed, and for the revalidation program proposed, the clerical and data input staff required would be considerable. One solution would have been to compete for these extra personnel resources with other components of the CF Personnel System, but research units tend to be at a competitive disadvantage when the economy is tight. The approach we have adopted is an extension of the solution described earlier in this paper. The CFTS routinely records on it computer files data on each trainee as to whether he successfully completed training, and if not, a disposition action and reason. This corresponds to the "Disposition" variable described earlier. An effective monitoring system is already in place at CFTS to maintain the quality of this data. CFPARU has for some time been the unit responsible for machine-scoring and storing recruit test scores. Here too, an effective monitoring mechanism is already in place. Arrangements have been made to obtain, on a continuing basis, copies of the appropriate CFTS computer tapes, and to use these to generate our validation files by merging these data with our own test score files. These steps ensure that our validation files contain a complete listing of all recruits who began training, as well as interpretable training performance data on those trainees who were unsuccessful. However, additional, more precise information would still be required on those trainees who were successful. This corresponds to breaking up the pass category using grade information to construct the consolidated criterion described earlier. This is to be accomplished by arranging for the routine transmission to CFPARU of training performance data from each of the trades training schools on all successful trainees at the end of each course. This data will be used to upgrade the CFPARU validation files. Thus, all the elements are at hand to construct a criterion variable which includes relatively precise performance data on all trainees. We are confident that, with this data, we will be able to improve substantially on our selection model, and extract the maximum predictive utility from the battery.

References

Rampton, G.M., Skinner, H.A., and Keates, W.E., Selection and Trade Assignment (Men) Project - Status Report. CFPARU Technical Report 72-7, Toronto, 1972.

MULTIVARIATE ANALYTIC STRATEGIES FOR VALIDATION OF "A" SCHOOL SELECTION COMPOSITES

Robert L. Frey, Jr.
U. S. Coast Guard Headquarters
Washington, D.C.

Robert P. Palese
U. S. Coast Guard Training Center
Governors Island, N. Y.

BACKGROUND AND PURPOSE

At this time, the United States Coast Guard (USCG) uses the Navy Basic Test Battery (BTB): General Classification Test (GCT), Arithmetic Test (ARI), Mechanical Test (MECH), Clerical Test (CLER), and Electronics Technician Selection Test (ETST) for classification and entrance to "A" school. In June, 1976, many of the selector composites for USCG "A" schools were lowered by a large amount. For example, the selector composite for Electronics Technician (ET) school had been GCT + ARI + ETST $\geq$ 170. The new cutoff score became GCT + ARI + ETST $\geq$ 155. With a waiver, the cutoff score goes down to 150. In percentile terms, the old Navy Standard Score (NSS) of 170 is approximately equal to the 75th percentile and the new NSS of 155 is approximately equal to the 56th percentile for an unrestricted mobilization population. The implementation of these lowered cutoff scores had the effect of giving much greater opportunity to minorities for entering an "A" school.

Since that time, the Uniform Guidelines on Employee Selection Procedures (43 FR 38290 et seq., August 25, 1978; 44 FR 11996 et seq., March 2, 1979) have been put into effect in final form. It is beyond the intent of this paper to present an overall review of the "Uniform Guidelines." Only those aspects of the "Uniform Guidelines" relevant to this validation study will be mentioned briefly.

The "Uniform Guidelines" apply, amongst others, to the Federal Government itself, state and local governments, most private employers, and, of course, to contractors and subcontractors of the Federal Government. The "Uniform Guidelines" define what practices are covered and under what conditions validation studies are required. Selection for training is a practice covered by the "Uniform Guidelines." Minimum technical standards for criterion-related, content and construct validity studies are also provided. In the case of criterion-related validity studies (such as this one) issues that are addressed include appropriate criterion measures, degree of statistical relationships, and test fairness. The distinction between differential validity and differential prediction (within the context of test fairness) is also discussed. However, the discussion in the "Uniform Guidelines" is typically at a fairly general level. There still is quite a degree of discretion in the specific mechanics of a criterion-related validity study.

Three years have elapsed since the USCG implemented the new, lowered cutoff scores for "A" schools. It seems appropriate to conduct a criterion-related validity study of the relationship between BTB scores and performance in "A" school under the new conditions. Specific objectives of the study included investigation for: 1) the statistical relationship between BTB scores and "A" school performance, 2) differential validity between ethnic groups, 3) differential prediction, 4) fairness, and 5) possible need for new selection composites. In doing the above, an attempt was made also to be in compliance with the "Uniform Guidelines" technical standards for criterion-related validity studies.

METHOD

The USCG "A" schools provide non-rated enlisted personnel with apprentice training in ratings. Everyone who graduates receives a final school grade which is derived from a combination of written exams and performance exams. A final school grade theoretically may range from a minimum of 70 to a maximum of 100. Those who drop out of "A" school do not receive any kind of "assigned" final grade.

Sample: Data were collected on the personnel who went to the five USCG "A" schools at Governors Island, N. Y. during FY 1978. The largest number of persons attended ET school. Accordingly, the ET students were selected as the sample for this study. The sample was composed of 354 whites and 30 blacks. (Other minorities had attended ET school but there were too few for meaningful analyses.)

Variables: The predictors included the BTB scores: GCT, ARI, MECH, CLER, ETST. The composite of GCT + ARI + ETST was also included since this is the current ET selection composite. The criteria were final school grade (FSG) and the categorical variable -- graduated vs. disenrolled (non-medical).

Data Analyses: Validity coefficients and summary statistics on all the variables were computed for whites and blacks separately. Blacks and whites then were compared on validity coefficients, regression lines (error of estimate, slopes, intercepts) and attrition rates. Also a crossed factorial multivariate analysis of variance (MANOVA) was done. The two factors were: 1) Ethnic Group - white vs. black, and 2) ET school performance. The latter factor was defined as a four level factor -- disenrollment vs. FSG in the 70's vs. FSG in the 80's vs. FSG in the 90's. The BTB scores were the dependent variables for the MANOVA. Discriminant function vectors were also obtained as part of the MANOVA approach.

RESULTS AND DISCUSSION

The blacks had a mean of 167.3 on the selection composite and the whites had a mean of 181.3. The difference between the two groups was significant ($p < .001$). The blacks had a mean of 86.1 on the FSG and the whites had a mean of 87.9. The difference between the two groups was significant ($p < .047$). The validity coefficient for blacks was .485 ($N = 22$ graduates, $p < .02$) and .501 for whites ($N = 300$ graduates, $p < .0001$). The test for differential validity showed that there was no difference in validity between the two groups. The attrition rate was 26.7% for blacks and 15.5% for whites. These percentages were not significantly different by the Chi-Square test.

An ANOVA approach was used to test for differential prediction.[1] Ethnic group was the factor -- blacks vs. whites. First an analysis of covariance was run separately on each group. When an analysis of covariance is done on each group, the MS error is the error of estimate for the regression equation. The FSG mean adjusted for the selection composite as covariate is the intercept. The regression coefficient is part of the program output as well. Comparison of the errors of estimate by Chi-Square showed no significant difference. Then an analysis of covariance was run on the two groups together. The test for equality of regression showed that there was no significant difference in slopes between the two groups ($p < .871$). The comparison between the FSG means adjusted for the selection composite as covariate is the test for differential intercepts. The test showed no difference in intercepts ($p < .961$). All results support the validity of using the same prediction equation for both blacks and whites in ET school. The raw score equation is:

$$\text{FSG} = 61.26 + .146(\text{selection composite score})$$

It should be noted that .146 is the within groups regression coefficient. This removes the confounding of mean group differences on the predictor and criterion.

The analyses done so far pretty well exhaust the so-called "regression model" approach to test fairness but that is only one of a number of competing models (see Petersen and Novick, 1976). The "Uniform Guidelines" do not support any particular model of test fairness but mention tests for differential validity and differential prediction as one possible approach. The differential prediction test is particularly needed since a test can be unfair even when differential validity is not found. Accordingly, the ET selection composite seems to be a fair test for both blacks and whites under the "regression model."

[1] All ANOVA, MANOVA and discriminant function analyses were accomplished using the MANOVA computer program written by Elliot Cramer. The program is available from the author: Dr. Elliot Cramer, Psychometric Laboratory, University of North Carolina, Chapel Hill, N. C. 27514.

However, the lack of obtained FSG's below 74 and the lack of grades for the dropouts seems to have distorted the regression equation, especially the intercept. The minimum passing grade is 70. A predicted FSG of 70 requires a selection composite score of only 59.9. This is utter nonsense, of course. The chance level score on the selection composite is 117! Assigning a failing grade of 65 to all the dropouts does produce a prediction equation which seems to make more sense. A predicted FSG of 70 would require a selection composite score of 130. Such a technique is completely arbitrary, though, since the intercept can be changed quite dramatically by one's choice of a grade for the dropouts.

Clearly, the dropouts must be included in the analysis as part of a valid experimental design which will simultaneously address the central issue of fair prediction of "A" school performance. Also, the straight line regression equation apprcach potentially can mask much meaningful information. For example, a crucial question is whether the current selection composite (or a composite with alternate weights) can be used to differentiate the dropouts from the marginal graduates, (i.e., those with an FSG in the 70's).

In order to address the above issues, an alternate experimental design was used. The design was a crossed factorial ANOVA with two factors: 1) Ethnic group (two levels) -- blacks vs. whites, and 2) School performance (four levels) -- dropouts vs. FSG (70's) vs. FSG (80's) vs FSG (90's). In this analysis the dependent variable was the selection composite score. The following is the experimental design with the selection composite cell means and cell N's in parentheses.

		70's	80's	90's
	Dropouts	FSG	FSG	FSG
Blacks	158.9 (8)	152.5 (2)	170.5 (15)	177.0 (5)
Whites	170.0 (54)	172.8 (10)	178.9 (14)	192.1 (106)

For the sake of brevity, the ethnic group factor will be called the G factor and the school performance factor will be called the P factor. In ANOVA terms, then, there are significance tests for G, P, and the G X P interaction.

The logic of the analysis then was: 1) Can the school performance categories be differentiated in terms of their selection composite scores, and 2) is the metric of the differentiation the same (or different) for the blacks and whites?

In terms of the ANOVA design, then, question 1 is the test of the P factor and question 2 is the test of the G X P interaction.

As is obvious from the cell N's, the design is non-orthogonal. That is, the SS due to G, P, and G X P are confounded. Therefore, a least squares hierarchial, step-down analysis was used. First, the G X P test was done eliminating the confounding of both G and P. Then the G test was made eliminating the confounding of P only and the P test was made eliminating the confounding of G only. The rationale is that the interaction test should be done first. If the interaction is significant, tests of main effects are not appropriate. When the interaction term is not significant, then tests of main effects are done ignoring the interaction SS.

The overall G X P interaction test was not significant ($p < .630$). The P factor test was significant ($p < .001$) and the G f.ctor test was significant ($p < .001$). When the dropouts are included in the analysis, then, there still is a strong relationship between selection composite scores and ET school performance. Further, this relationship holds above and beyond the confounding effects of ethnic group differences on the selection composite. The lack of a G X P interaction shows that the relationship between selection composite scores and school performance is the same for blacks and whites. In terms of "Uniform Guideline" requirements, there is still a valid relationship between selector and criterion and the selector is fair for both blacks and whites.

However, a crucial question for determination of a cutoff score is whether or not the selection composite can differentiate the dropouts from the marginal graduates (70's FSG). A planned contrast between the dropouts and the 70's FSG group was not significant ($p < .770$). (As before, the test removes the ethnic group confounding.) The interaction between ethnic group and this planned contrast was not significant ($p < .431$). This means that for both blacks and whites the dropouts cannot be differentiated from the marginal graduates on the selection composite. Additional planned contrasts showed that the 80's FSG group has significantly higher selection composite scores than the dropouts and 70's FSG group ($p < .001$), and that the 90's FSG group has significantly higher selection composite scores than all the others ($p < .001$). The interactions of these contrasts with ethnic group were not significant. That is, the level of differences between FSG groups are the same for both blacks and whites.

What does all of this mean in terms of determining cutoff scores? Before we deal with that issue, some overall discussion of the underlying logic of the experimental design is necessary. The "regression model" approach used earlier had determined an equation for predicting school performance given an individual's selection composite score. The error of prediction is on the criterion side. That is, a predicted school performance score is theoretically the mean score for all of the people with that same selection composite score.

The crossed factorial design analysis in effect reversed the logic. People were first grouped into four school performance categories: dropouts vs. 70's FSG vs. 80's FSG vs. 90's FSG. Then the question becomes -- given the school performance categories, can these categories be differentiated in terms of their average selection composite scores?

If these categories can so be differentiated, then meaningful prediction of school performance is possible. This logic is, of course, the same line of reasoning used in discriminant function analysis. In this case, there was only one predictor score instead of a set of predictors. (The final analysis of this study will indeed be a discriminant funtion analysis using the individual GCT, ARI, ETST scores.)

Using the same approach to classification as is done with dis^riminant function scores, we will now present some possibilities for setting cutoff scores. Based on the results of the significance tests, we now assume there are 6 groups of people in terms of their selection composite means. The dropouts and 70's FSG category are one group in terms of their selection composite means. Thus, there are three discriminable school performance categories. Further, the G X P tests were not significant and the G test was significant after the confounding of the P effect was removed. These two findings together tell us that the blacks and whites are significantly different on selection composite means by the same amount within each school performance category. In ANOVA terms, then the significance test results tell us that we have two ethnic groups X three school performance categories with no interaction.
For cutoff score purposes we will now redo the design as established by the significance tests with appropriate selection composite means:

	Dropouts & 70's FSG	80's FSG	90's FSG
Blacks	159.7	168.1	180.9
Whites	170.7	179.1	191.9

(These cell means are reconstructed from the least squares marginal means. Such estimates are more accurate than the original cell means in a model without interaction.)

First of all, the blacks and whites are two separate populations in terms of the average selection composite score needed to achieve a given school performance category. Therefore, blacks and whites are treated separately for the purpose of setting a cutoff score. Secondly, the dropouts cannot be distinguished from the marginal graduates. This makes the cut score selection process more complex. If the dropouts and 70's FSG group were differentiated in terms of their selection composite scores, the process would have been more straightforward. Within each ethnic group, the 70's FSG category mean on the selection composite would have been noted. This selection composite mean would be the "group centroid" for the 70's FSG category. Then a confidence interval would be put around this group mean. A simple approach would be halfway between the dropout mean and the 70's FSG category mean. When a new applicant would apply for ET school, his/her score would be compared against the minimum score point on the confidence interval. Often the length of the confidence interval (e.g., in standard error units) becomes a matter of judgment in terms of organizational policy. To note the obvious, the wider the confidence interval, the more opportunity for people with lower selection composite scores to qualify for a given "A" school.

Now let us return to the more complex situation with the present results. The data do not support a cutoff score determined from the 70's FSG category mean. A number of possibilities present themselves. We will discuss the most obvious ones. Some organizations might want to minimize not only dropouts but the marginal graduates as well. In this case, the cutoff score might be halfway between the 80's FSG category mean and the dropout/70's FSG category mean for each ethnic group. For the blacks, the cutoff score would become (159.7 + 168.1) / 2 = 163.9 = 164. For the whites, the cutoff score would become (170.7 + 179.1) / 2 = 174.9 = 175. Let us hasten to add that we have no intention of recommending such scores but went through the exercise for illustrative purposes.

If we focus on the dropout/70's FSG category mean to determine a cutoff score a number of judgments have to be made. Looking at the blacks in this category, only 20% of them graduated. Looking at the whites in this category, only 15.6% of them graduated. It seems that for both ethnic categories, these people are a high-risk group. One possibility, then, would be to use the group means without a confidence interval. That is the black cutoff score would be 159.7 = 160 and the white cutoff score would be 170.7 = 171.

Again, we are not recommending these scores. Organizational policy considerations would have to be brought into the decision process. Whatever the policy decision, however, we feel that these group means are an objective starting point justified by a validation procedure. The cutoff scores can be extended downward (or upward) in standard error units to accomplish tradeoffs between such factors as attrition rates and providing opportunity for entrance into "A" schools. However, establishing different cutoff scores for blacks and whites has potential for charges of reverse discrimination. We definitely feel that legal consultation would be necessary before attempting such a policy. The statistical justification, of course, is that each group has the same predicted FSG.

As a methodological note, however, we should remember that the lower cutoff score for blacks was not predetermined from the fact that the black mean on the selection composite was so much lower than the white mean (167.3 vs. 181.3). When black and white means are compared within non-dropout FSG categories, it would be entirely possible for the result to be equal means (and thus equal cutoff scores). Theoretically, such a comparison could even demonstrate higher black means for the non-dropouts within FSG categories. This result would call for higher cutoff scores for the blacks. In other words, the experimental design approach used in this study can possibly produce the same range of ethnic group cutoff score combinations as with the regression equation approach.

The last analysis to be discussed used the same experimental design as before (G X P). In this analysis, however, the dependent variables were the GCT, ARI, and ETST scores. In the previous analysis, by definition, the GCT, ARI, and ETST scores had equal weights when their composite was used. The purpose of this analysis is to find the linear combination of GCT, ARI, ETST scores that best differentiates the school performance categories. Essentially, this is a discriminant function analysis incorporated into a crossed factorial design. The G X P interaction was not significant (p < .804). The P test was significant (p < .001, R = .515) and the G test was significant (p < .001, R = .256).

The P test standardized discriminant function coefficients were .066, .623, and .518 for GCT, ARI, and ETST respectively. These coefficients indicate the relative contribution of each variable to differentiating the school performance categories. By simple inspection, we see that the ARI score and the ETST score are making the predominant contribution to differentiation. The GCT score appears to be making virtually no contribution to differentiating the school performance categories. If this result were to hold up under replication and larger minority N's, it would point to the possibility of needing only the ARI and ETST composite for ET school selection.

Analysis of covariance is sometimes used to decide whether a given variable is making a significant relative contribution to differentiation. In this particular case, GCT would be the single dependent variable and ARI and ETST would be the two covariates. If the P test on GCT with ARI and ETST as covariates is significant, this would indicate that GCT is making a contribution and should be retained in the selection composite. Such a test was made. The P test on GCT with ARI and ETST as covariates was not significant (p < .898). This implies that GCT is not making any contribution to the differentiation of FSG categories.

The planned contrasts showed that the dropouts and the 70's FSG category could not be differentiated even with the optimal weights of discriminant function analysis (p < .781). No interactions were significant. The remaining planned contrasts showed that the 80's FSG category and the 90's FSG category were distinct groups. P factor discriminant function scores were computed and then ANOVA was applied to these scores to obtain the least squares estimates of the cell means reconstructed from the marginal means. This resulted in the following table of discriminant function score means:

	Dropouts + 70's FSG	80's FSG	90's FSG
Blacks	11.0	11.7	12.6
Whites	11.6	12.3	13.2

When using discriminant function scores to set cutoffs, the logic of the procedure is the same as before. Policy and judgment considerations still remain of course. To give an arbitrary example for illustration purposes, let us assume a cutoff score of 11.0 for blacks (i.e., the black mean for the dropouts/70's FSG category). The raw score discriminant weights are applied to the GCT, ARI, and ETST scores. These are .01141, .11215, and .08707 respectively. Then, the resultant score is compared with the cutoff of 11.0. Of course, many different combinations of GCT, ARI, and ETST scores can produce a discriminant score of 11.0. The same principle operates in producing any given value of the simple sum composite as well. To give a feel for the average BTB score needed to produce each discriminant score, the following table applies. The main entries are the discriminant function scores with the average score required on GCT, ARI, and ETST in parentheses.

	Dropouts + 70's FSG	80's FSG	90's FSG
Blacks	11.0 (52.2)	11.7 (55.5)	12.6 (59.8)
Whites	11.6 (55.1)	12.3 (58.4	13.2 (62.9)

That is, if a person had a GCT score of 52.2, an ARI score of 52.2, and an ETST score of 52.2, the resulting discriminant function score would be 11.0. The reason for going through all this trouble of course is that the discriminant function approach provides the highest possible accuracy for classification of people into predicted FSG categories. It should be noted that discriminant function weights do have the same problem as multiple regression weights. They are volatile across samples and cross-validation is imperative before recommending new composites.

In summary, this study applied both the regression model approach and the crossed factorial discriminant function approach to the validation of a selection composite. Apparently because of the lower end cutoff on FSG scores, the regression prediction approach could not be used for determining cutoff scores. The discriminant function approach did seem to provide valid starting points for making cutoff score decisions. Furthermore, this approach allows for determination of selector fairness by means of Ethnic group X Criterion category interaction tests. Of course, one study does not conclusively prove the superiority of any technique. Nevertheless, we do feel that the potential for discriminant function analysis as a criterion-related validation technique certainly should be explored further.

REFERENCE[1]

Peterson, N. S. & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

[1] The entire issue of Journal of Educational Measurement, Volume 13, Number 1 (Spring 1976) is devoted to the selection bias problem.

VALIDITY CONSIDERATIONS IN THE DESIGN OF MANAGEMENT SURVEY INSTRUMENTS

John F. McAreavy, Ph. D.

U.S. Army Management Engineering Training Activity
Rock Island, Illinois 61299

Introduction

Managers and other decision makers faced with high order complexities and interfaces are becoming more and more dependent on studies, surveys and audits to provide information necessary to make decisions on base closings, consolidations, and internal reorganizations. In the government setting, these activities are referred to as management studies, staff studies, management audits, program evaluations, operations audits, management surveys, pre-award surveys, manpower studies, etc. Due to the significant, far-reaching consequences, it is important that decision makers and analysts fully consider the validity of the data and procedures used in conducting these studies.

If the organization under study is characterized by routine production activities or precise laboratory procedures, the operational measurements obtained for subsequent analysis can be validated by classical procedures. ['owever, if the organization is being studied at the macro level where the emphasis is on the goals, resource utilization and productivity, the measurements tend to be less precise and, in many cases, the result of individual opinions and subjective observations of individuals and groups. In these cases, validity checks are difficult to perform, so analysts generally make a cursory caveat about "inherent" validity and quickly proceed to the analysis portion of the study. This paper describes a technique that enables one to enhance the validity of a management audit, but could be used in all management study areas.

The Management Audit

The management audit, as it is currently evolving, varies from other industrial or financial surveys conducted by governmental or business organizations in that it focuses upon performance or management accomplishment rather than solely upon procedures, processes, and organizations. In this sense, the management survey is a methodical examination of how management is getting things done. It is a measure of management accomplishment.

The worth of any management audit report is wholly dependent on the validity of the specific process and criteria utilized in performing the audit. Since the objective is to measure management capability in an organization, the analyst must have assurance that the audit does, in fact, measure that capability. Two factors are involved at this point: (1) The design of the management audit structure, content, and criteria, and (2) the conduct of the audit. The overall validity of the audit is influenced by both of these factors. One could have a highly valid audit design, but if the audit was poorly conducted, the validity of the audit report would be diminished. On the other hand, if the audit design had no validity, the audit report would also possess no validity.

Management Audit Objectives

The first and most important step in conducting a valid management audit is to identify the purpose and define the objectives of the audit. The basic purpose of the management audit is to carry out certain investigations and examinations of an organization to be able to predict with reasonable accuracy the success or failure of future management activities. From this purpose, three basic objectives of management audit emerge:

- To evaluate organization capability and performance to determine levels of efficiency and economy;

- To measure accomplishment versus programs and plans to determine effectiveness; and

- To identify operating deficiencies and potential improvements.

In addition to these basic objectives, each audit would possess other more specific objectives that would portray the audit's unique aspects. The basic and specific objectives would constitute the framework by which the content of the audit would be developed. The subsequent procedures were developed in order to conduct a management audit with the three objectives stated above, along with the specific objective of differentiating among the management capabilities of three large corporations who were competing for the large-scale development contract for a major Army weapons system.

Components of Design Validity

Design validity has two components: (1) the validity associated with the degree by which the audit is representative of the entire organization (company or corporation) and with its capability to ascertain true effectiveness and efficiency of the organization and, (2) the validity associated with the development of the questions, checklists, and other data collection instruments. In order to address the first component of audit design validity, one can adapt the process used by test designers and classroom teachers in developing test plans to assure content validity of achievement tests.

To serve the achievement testing purpose, the test designer usually prepares a plan with two dimensions; one dimension for content outline and the other for process objectives. This mechanism assures the designer that the test will be comprehensive and possess high content validity.

The Management Audit Matrix

Just as the test designer is concerned with content validity, the management audit designer must be assured that the audit is comprehensive and penetrating. This assurance can be effectively realized by utilizing a management audit matrix where the X axis consists of the functional categories that will make up the total organization under audit, and the Y axis consists of those management performance characteristics that affect the overall effectiveness and efficiency of the organization. A cell formed by the intersection of the X and Y axes contain the required benchmarks that specify minimum standards of management performance for that particular functional area and the corresponding performance indicator.

For an audit designed to assess the management capabilities of a large weapons systems contractor, the X axis could consist of five areas:

Executive
Project
Engineering
Manufacturing
Business

These five areas represent major organizational focal points existing throughout most industries. These areas and a brief description of each follow:

Executive Area: Decision and policy area of the organization, including the chief executive, his functional staff, and directors of the line organizations. This is the area where the auditors would expect to find the "prime movers" of the organization.

Project Area: That portion of the organization in which projects are managed. This includes interfaces with all functional activities of the company involved with the project(s).

Engineering Area: The management of the total engineering process of the organization to include technical planning and control, the decision making processes, and communication links with other functional elements.

Manufacturing Area: This area includes all aspects of manufacturing/ production and associated operations such as production control and material operations.

Business Area: This area has been selected because, in order to do a thorough evaluation, it is necessary to include those functions that support Project, Engineering, and Manufacturing, and tend to be part of the corporate staff rather than the line structure. Regardless of organizational placement, the business functions performed relate indirectly to the success or failure of the organization. The sub-areas in this area could well be automatic data processing, finance and accounting, purchasing and procurement, and contract administration.

The Y axis of the matrix consists of categories of management performance characteristics such as organization structure, personnel, responsibilities and authorities, and requirements responsiveness. These are the characteristics that shape the effectiveness and efficiency of the organization. Brief descriptions of these characteristics are as follows:

Organization Structure: Organization philosophy, planning effectiveness and interfaces.

Personnel: Personnel management practices, capabilities of the managers and key technical personnel.

Responsibilities and Authorities: Assignments of functional and management responsibilities, charters, decision making authority, problem resolution chains.

Requirements Responsiveness: The management and conduct of operations in response to customer requirements. Effectiveness and efficiency.

The matrix is portrayed below as Figure 1.

MANAGEMENT AUDIT MATRIX

Management Performance Characteristics (Y)	(X) AREA				
	Executive	Project	Engineering	Manufacturing	Business
Organization Structure		X_2Y_1			
Personnel	X_1Y_2				
Responsibilities & Authorities					
Requirements Responsiveness			X_3Y_4		

Figure 1

Benchmarks

At the matrix intersections (cells) of the X axis and Y axis, benchmarks that specify desired traits and performance criteria pertinent to this cell are developed for the auditors. Examples of benchmarks by matrix cell are illustrated as follows:

Matrix Cell X_2Y_1

(Project-Organization Structure)

- Establishment of a formally structured and documented project organization.

- Existence of a basic organizational philosophy, implementation rationale and control of project functions and resources.

Matrix Cell X_3Y_4

(Engineering-Requirements Responsiveness)

- Estimating within engineering should be compatible with the budgeting structure to ensure traceability from estimates to budgets to actual performance.

- Engineering controls must be established in the design process to prevent either under or over engineering and provide an efficient means for producing a quality design meeting customer requirements.

Matrix Cell X_1Y_2

(Executive-Personnel)

- Existence of specific criteria addressing the company's executive promotional and selection methods.

- Development and implementation of key executive performance requirements.

- Specific criteria addressing tangible methods of determining quality of performance.

The number of benchmarks in each cell is determined by analyzing the objectives and the organization's structure in light of the content of the specific cell. A cell will have more benchmarks when

(1) the subject content of the cell is dominant in one or more survey objectives;

(2) a large measure of the organization's structure or resources are associated with the cell.

In a recent application of this matrix technique, only three benchmarks were required for the X_2Y_3 cell (Project-Responsibilities and Authorities), while 53 benchmarks were required for the X_5Y_4 cell (Business-Requirements Responsiveness). In this application, 337 benchmarks were needed in the entire matrix to achieve the stated objectives of the audit.

By developing this matrix of benchmarks, the first component of design validity is fully considered. The second component of design validity consists of the evaluation process that is developed to ascertain the degree of compliance with the benchmarks.

Statements, Questions and Checklists

The degree of compliance to the benchmarks is determined by probing the organization's operations using definitive statements, questions and checklists (SQC). These SQC must provide sufficient documentation for the auditor in order that he may objectively assess the degree to which the benchmark is attained. Whereas the benchmark is a concise baseline of acceptability of management practice, the SQC must provide relevant detailed descriptions of the benchmark. In order to attain maximum validity, the SQC must provide complete and firm definitions of the quantity of acceptability, quality of acceptability and minimum conditions of acceptability. The SQC must have the necessary characteristics for use in examination and audit at each level in the organization. Also, the SQC must possess the capability to validate and cross check the audit information and must be designed for interview use or for adaptation to other techniques of audit such as work sampling, document review and evaluator observation. Instructions for application must be prepared for each series of benchmarks and the SQC must provide for summarization procedures and a guide for interpretation of results. The number of SQC will vary among the matrix cells as their density is a function of the emphasis, impact or complexity of each benchmark.

Summary

This process, generating a matrix of benchmarks and a matrix of statements, questions, and checklists provides assurance that the management audit is designed with maximum content validity. Competent,experienced management auditors using this design will then be in a better position to provide an audit report that can be used with confidence by managers who are faced with making significant decisions.

BIBLIOGRAPHY

Books

Cronbach, Lee J. Educational Psychology. Harcourt, Brace and Company, 1954, New York.

Lindquist, E. F. (Ed.). Educational Measurement. American Council on Education, 1951, Washington, D.C.

Norbeck, Edward F. et al. Operational Auditing for Management Control. American Management Association, Inc., 1969, New York.

Thorndike, Robert L. and Hagen, Elizabeth. Measurement and Evaluation in Psychology and Education. John Wiley & Sons, Inc., 1955, New York.

Other

Air Force Systems Command Pamphlet 70-1, "A Summary Of Lessons Learned From Air Force Contractor Procurement Systems Review," 1965.

Armed Services Procurement Regulation Supplement No. 1, "Guide For Conducting Contractor Procurement System Review (CPSR), 1966.

Faucett, Phillip M. "Management Audit for Small Manufacturers," Small Business Management Series #29, Small Business Administration, Washington, D.C., 1963.

McAreavy, John F. "An Analysis of Factors Affecting the Achievement of Adults Who Participate in Short Concentrated Courses," Dissertation, U.S. Army Management Engineering Training Activity, Rock Island, IL, 1969.

Worthy, James C. "The Management Audit, " Proceedings of the Annual Meeting, Academy of Management, 1962, p. 176.

PERFORMANCE EVALUATION OF ENLISTED WOMEN DURING REFORGER 77

Bertha H. Cory and Cecil D. Johnson

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

PROBLEM

Since 1972, considerable attention has been directed toward determining the impact of expanding the role of women in the Army. Prior to 1975, the relatively small number of women in the Army were, with rare exception, placed in occupations that are, in our society, traditional for women. As increasing numbers of women were recruited, it quickly became necessary to use the additional numbers in more non-traditional roles, mostly in combat service support units. The increasing number of women being assigned to these units raised questions as to the impact of the presence of women on the Army's future effectiveness on the battlefield. In 1975, the Army Research Institute was asked to provide data for objective conclusions as to whether the performance of support units would be impaired when the units deployed into the field with prescribed numbers of women. The prescribed percentages of women corresponded to unit quotas for women, from 10% to 20% for divisional units, and 35% for units in the theatre but further from the front lines.

Although Congress has provided guidance that women should not be assigned to combat roles, women may be assigned to support units such as Military Police, Signal, Maintenance, Transportation, and Medical that have the mission of providing support to maneuver battalions (Infantry, Armor, etc.). Such support units can be called upon to defend unit perimeters and perform other contingency missions against an enemy force. The question had to be answered: can women soldiers comprise a sizable proportion of Army support units without impairing the unit's capability for accomplishing battlefield missions?

BACKGROUND

The Army Research Institute for the Behavioral and Social Sciences (ARI) conducted a force development test "Maximum Women Army Content," referred to as "MAX-WAC," to determine the effect of different percentages of enlisted women soldiers (EW) on unit performance in a 72-hour exercise. A total of 55 company-sized field exercises were observed at posts within the continental United States. Of these, 25 companies were in only one scored exercise each, and 15 companies were each in two scored exercises 6 months apart. In the second exercise, personnel were experimentally shifted so that 5 companies contained 0% to 15% EW and 5 companies contained 15% to 35% EW. The major finding of MAX-WAC, as related in the final report of October 1977, was as follows: On the average, the 0% to 15% shift resulted in a slight decrease in performance scores while the 15% to 35% shift provided a slight increase in performance scores. Neither shift resulted in differences that were statistically significant since the two different directions (the non-linear relationship between number of women and performance) had not been hypothesized. The ARI staff predicted that a repetition of the MAX-WAC experiment with more companies, improved instrumentation, and better control of extraneous factois would yield essentially the same conclusion: no significant difference.

The MAX-WAC conclusion that women soldiers could live and work effectively for three days under field conditions did not necessarily permit extrapolation to the actual battlefield. One of the major disconnects, as perceived by several field commanders, was the short time span of the MAX-WAC exercises. The question was seriously posed as to whether women soldiers could do as well over extended periods in the field. The investigation now being reported was planned in the spring of 1977 to address this question by taking advantage of a major field exercise scheduled to occur in the fall of 1977 called REFORGER 77 (Return of Forces to Germany). The ARI investigation was titled REF-WAC, as an abbreviation for REFORGER-Women Army Content.

APPROACH

The annual REFORGER exercises in Germany involved one and a half weeks of realistic war games with division-sized forces on each of two opposing sides. The U.S. Forces on one side were transported from one or two installations in the U.S., while the other U.S. Forces were those already stationed in Germany. The soldiers from the U.S. were absent from their home installations for about six weeks, three weeks of which were under field conditions in Germany.

Since some support units on both sides of the REFORGER 77 exercise contained close to 10% women, it was considered feasible to follow closely the performance of enlisted women (EW) and counterpart enlisted men (EM) matched on demographic and personal characteristics. Other comparisons between the matched men and women soldiers related to deployability and to time lost from duty.

REF-WAC observer teams collected data using standardized performance criteria in the following five divisional combat support and combat service support units: military police, signal, medical, maintenance, and supply and transportation.

Small teams or work groups with a sizable number of women were compared with similar all-male units performing the same standard tasks in both the early and the later parts of REFORGER 77. In this manner, the Army Research Institute addressed the question of whether group performance over an extended field exercise was affected more adversely by fatigue and stress in units containing women than in those with all male personnel.

Women soldiers were matched with male cohorts on the basis of similar service time, grade, and military occupational specialty. Standard tasks scorable for individual performance were used to compare women and male cohorts during the REFORGER. Both the standard tasks designed for teams and for individuals were scored by officers assigned to a REF-WAC Test Directorate.

All women soldiers and the selected male cohorts were also rated daily as to overall performance by non-commissioned officers (NCOs) who were serving as their work-place supervisors, and, as appropriate, by non-commissioned officers supervising them in special duty status (as in guard duty). Both soldiers and unit supervisors filled out questionnaires just before and immediately after the field exercise. The supervisors were also asked to specify the positions to which they would assign women, for different proportions of women assigned. Each supervisor was given 5 sheets listing the exact MOS with authorized numbers for his unit's organizational table. On the first page, he was to give his preferred distribution of men and women assuming 10% women, on the second assuming 35% women, the third 50%, the fourth 65%, and the fifth 90%.

FINDINGS

One of the more interesting conclusions was that the performance of neither men nor women was impaired by remaining in the field in an overseas exercise longer than the 3 days utilized in MAX-WAC. Indeed, when daily performance ratings by supervisors in high stress companies were considered separately, enlisted women gave poorer performance (statistically significant) than enlisted men initially (i.e., during the first 3 days), but gained complete equality in performance by the last 3 days of the exercise.

More generally, aggregated individual and group performance data for both men and women showed a clear upward trend over time from the first 3 days to the last 4 days of the exercise. Thus, REF-WAC results provided a basis for increased credibility of findings in the earlier MAX-WAC research that had been criticized because of the short time of its field exercises.

Judging from the questionnaire results, NCOs rated EW in the abstract lower than they rated specific women. This finding suggests that an interesting sociological phenomenon which often occurs with respect to minority group members also occurred in REFORGER. The individual members of a minority are accorded their observed value but the unobserved members of the minority group are presumed inferior.

With respect to questionnaire items on how well EW performed on REFORGER 77, EM were the most critical, NCOs next most critical, and officers least critical, except for EW who rated themselves as highly as EM rated themselves. Two interpretations are possible: (1) proximity and presumed opportunity to observe produces lower ratings, except for women rating themselves highly; (2) EM who worked in units with EW were negative toward EW because of the tendency to assign men to more of the "extra" work, such as harder physical labor and night duty.

When supervisors were asked to hypothetically distribute EM and EW to MOS positions in the unit organizational table for five different proportions of EM to EW, three sometimes conflicting assignment policies emerged, described as "concentration," "proportionality," and "suitability."

(1) A comparatively small number of women were placed in the MOS believed to require the greatest strength and a comparatively large number were placed in the MOS believed to require the least strength. In contrast, men were placed in the "harder" MOS. This concentration of men and women according to the perceived physical difficulty of the MOS tasks was most evident when the hypothetical task was to place a small number of men or women in a company which was predominantly of the other gender.

(2) When the hypothetical task was to place larger percentages of either men or women against the company organizational positions, supervisors assigned women more broadly across MOS. Men were similarly spread out, probably to assure some men in each group to do the more physically demanding tasks. This tendency to place women or men in numbers proportional to the numbers needed in each MOS reduced the concentration of men or women in jobs seen as physically difficult or easy.

(3) In some maintenance and some supply and transportation companies the hypothetical assignment of 10% men (with 90% women) was apparently determined primarily by the frequency with which the MOS occurred in teams assigned to work forward of the brigade rear boundaries--in the midst of the combat arms units, called FAST and CONTACT. This third policy, suitability, resulted in a comparatively small number of men being available for placement in the more difficult physical tasks. It should be noted that REFORGER 77 occurred before dissemination of the new DA policy permitting female soldiers to be assigned anywhere on the battlefield. The new policy only prohibits the assignment of women to units whose primary function is to seek out and engage the enemy. At the time of REFORGER 77 supervisors still felt prohibited from permitting female soldiers to be assigned forward of the combat brigade rear boundaries as members of FAST and CONTACT teams. Thus, a number of factors important in shaping opinions on the most appropriate use of women may or may not still apply.

In a follow-up of some of the EW who had been on REFORGER there was no evidence that women who had been athletic or participated in outdoor activities such as camping performed better on REFORGER 77 than those who had not done this. The most important factors in the quality of the performance of EW on REFORGER 77 were training and experience on the jobs specifically required of them on the exercise. The poorer performers generally were those who were assigned tasks for which they had little or no preparation.

Contrary to opinions of many who extrapolated from their experience in combat and on all-male extended field exercises, women did sustain themselves in the field and accomplish MOS-related duties at an acceptable level. However, there was questionnaire evidence that women were less willing than men to play the role of a support unit soldier in the event that the battlefield was a real one.

METHODOLOGICAL CONSIDERATIONS

Most of the limitations of the REF-WAC research effort stemmed from the severe time constraints under which the research was planned and executed. The research was initiated only 6 weeks before the deployment to Germany of officer observers and NCO data collectors. During this interval all instruments were planned, constructed, and printed, together with procedures for their use. Major data analyses were completed and Special Report S-7 was written only 10 months from day one of the project. Clearance, printing, and distribution occurred within 1 month more.

One of the serious limitations produced by the time constraint was that officer-observers for the five types of units were not located at ARI long enough to acquire a thorough understanding of the rationale for the standardized performance criteria which they were to use for observations during REFORGER. In a few cases the members of the officer teams were identified and brought on board only a few days before deployment. Like many junior executives in industry, many of these officers had not had sufficient prior research orientation to appreciate the necessity of collecting standard data in carefully structured ways in order to draw scientific conclusions. Further, they made the error of over-concentrating on observing performance in units containing women and thus not giving equal attention to all-male units for comparison purposes.

Similarly, the NCO data collectors were not available to ARI scientists for systematic instruction in their duties. Also, the process of selecting these NCOs was rushed. Three considerations would have been useful in selection of the NCOs:

(1) higher than average ge.eral ability, (2) MOS experience similar to that of the EM and EW to whom they were to be assigned, and (3) an interview evaluation by ARI project scientists.

Without opportunity for tryout of the daily performance recording procedures, assessment of size of the NCO work load was not possible. It is clear now that most NCOs (those who did not have to cover widely dispersed units) had time to have collected a great deal more useful information such as: performance on specific MOS tasks, performance on key tactical and sustenance tasks, data on differential utilization of men and women.

All instruments--rating forms, questionnaires, instructions--had to be used without opportunity for tryout and refinement. Obviously numerous improvements would have resulted from tryouts.

At the time the decision was made to conduct this research, the structure and plans for REFORGER 77 field exercise had been determined. This fact created another restraint on ARI's research plans: that of non-interference in the war game. "Events" which were observed and or which performance of EM and EW was rated were entirely "targets of opportunity." Events for which standardized rating forms were created were chosen only upon the basis of expert judgment as to their probability of occurrence. There could be no standardization of content scenarios, no controlled variation in difficulty, frequency, or variety. As a result, activities which women soldiers performed and were evaluated on were most often not the tactical-type tasks. Many units had too light a work load, did not move, and were generally under little stress; they did not engage in perimeter defense or other similar jobs. Thus tactical tasks as well as many other critical MOS tasks were infrequently observed. There was no possibility of controlling the proportion of women in the units observed--which would have been useful to research design. Also non-controlled was the method of assigning EM and EW to "special duty" tasks; as a result it appeared that EM were given a greater share of night duty, physically demanding tasks, and tasks away from the unit location.

Other factors affecting findings and their interpretation include the following: (1) the weather generally was mild and dry; (2) clothing provided female soldiers was often inappropriate even for these mild conditions; (3) as far as could be determined, none of the women soldiers had had the basic training, which now is generally equivalent for EW and EM. The women soldiers were handicapped in the first days of the field exercise by insufficient experience with tactical and sustenance tasks--presumably now being more adequately taught to EW in basic training.

In any future REF-WAC-type research, the following recommendations should be considered: (1) The exercise scenarios should have the training of support companies as an important objective, rather than virtually ignoring support training requirements; this would cause personnel in these support units to perform more realistic combat-like tasks--tasks they do not perform when the scenarios are written solely for the combat arms units. (2) The fill within units should be controlled as to proportions of men and women soldiers; some sections could thus have 20% to 30% EW although the company still falls within the prescribed quota of 10% EW. (3) NCOs should be carefully selected and trained to collect performance data on tasks; this might possibly eliminate the need for the officer observers utilized in REF-WAC 77. (4) Post-exercise questionnaires should be shorter and focused on key issues; identification of selected EW, male cohorts, and supervisors would permit follow-up and interview after 3 months.

In conclusion, all REF-WAC findings must be examined with the recognition that any maneuver is a simulation which imperfectly predicts combat performance. This imperfection may be more extreme for support companies than for combat brigades.

REFERENCES

Army Research Institute. Women content in units force development test (MAXWAC). (ARI Special Report S-6). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1977.

Johnson, C. D., Cory, B. H., Day, R. W., and Oliver, L. W. Women content in the Army - REFORGER 77. (ARI Special Report S-7). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1978.

Oliver, L. W., and Babin, N. E. The relationship of gender to Army field assignment patterns. (Technical Paper). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1979.

Oliver, L. W., and Johnson, C. D. An investigation of the validity of performance scores adjusted for environmental conditions. (Technical Paper). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, in process.

THE AIR FORCE FEMALE PILOTS PROGRAM: AN INTERIM REPORT

Jeffrey E. Kantor and Capt Dana R. Ideen

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

INTRODUCTION

Although women have been involved in civil aviation since its inception and were utilized in a variety of non-combatant flying roles during World War II, at the conclusion of the war no continuing program was established and the female military pilot contingent was demobilized. With few exceptions, military aviation was a male domain. However, during the 1970's, the role of the military woman was expanded and in 1976 the Air Force opened undergraduate pilot training (UPT) to qualified candidates of both sexes.

Military aircraft and flying are quite specialized, demanding both a high degree of skill on the part of the aircrew and an optimum interface between the pilot and the aircraft. At the onset of the Air Force female pilots program there were many unanswered questions and concerns. Some issues were specific and safety-related, such as whether the women's size, reach, and strength were sufficient to fly current military aircraft. Other concerns were more general, such as whether selection and classification procedures developed and refined on an all-male population could be equally applicable and effective for women. Additionally, it was unknown how well women would learn and perform in a training environment that was optimized for an all-male population and represented a traditionally male role in the Air Force.

Since the beginning of the female pilot program, the Personnel Research Division, Air Force Human Resources Laboratory (AFHRL), in cooperation with other Air Force agencies, has been conducting research on the training, performance, and utilization of women from pilot training through their first few years of operational flying duties. Aptitude and attitudinal measures were taken prior to UPT while performance and attitudinal measures were later obtained at the conclusion of UPT, at the conclusion of advanced training, and at several points during the women's initial operational tours. The objectives of this research were to (a) establish a data base from female pilot candidates composed of paper-and-pencil, psychomotor, and aircraft simulator aptitude measures; (b) compare these data with those previously obtained from male subjects both for overall performance and for predictive efficiency, i.e., UPT training outcome; (c) monitor the attitudinal response of both the female pilots and male counterparts; and (d) monitor the flying performance of women as judged in terms of official Air Force flight standards and relative male performance.

This is the second in a series of reports documenting AFHRL research on the female pilots program. To summarize the first report (Kantor, Noble, Leisey, & McFarlane, 1979), few significant differences were found between men and women entering UPT. Comparable performance on most pre-training aptitude measures, combined with equivalent graduation rates, factors associated with flight training performance, and student impressions of the flight training experience, all lent strong support to the conclusion that men and women behave similarly in flight training. However, flight instructor ratings of male and female student characteristics did reveal several areas in which males were rated significantly better but the factors underlying these differential ratings wêre not discernible from the available data. Overall, the similarities between the sexes greatly outweighed the differences, which indicated that coeducational UPT could be accomplished without significant modification to the pilot training system.

While the previous report dealt with data from the first two groups of women to complete UPT, this report will utilize data from additional women who have completed UPT and will present preliminary information on the performance of women in advanced training and operational flying. It should be noted, however, that this interim report will not provide conclusive findings since more work remains, pending the collection of additional data from the women in Air Force flying careers.

METHOD

Subjects

Data were collected on a total of 30 female subjects who composed the population of women who had entered UPT as of July 1978. These women were selected from active duty and reserve components of the Air Force using the same selection criteria used for male candidates. From the 30 women who entered pilot training, 22 successfully completed UPT and these women became the subjects of evaluation during their advanced training and operational flying.

A total of 64 male subjects were used in this study. All male subjects were instructors for either UPT ($\underline{n}$ = 8) or one of the advanced phases of flight related training; Survival-Resistance training ($\underline{n}$ = 11), Replacement training ($\underline{n}$ = 39), or in-unit training ($\underline{n}$ = 6). All instructors had experience training both men and women.

Data Collection

Surveys were administered to the instructors to collect information on gender comparisons in the following six areas: (1) Attitude toward training, (2) background knowledge relevant to training, (3) ability to acquire skills, (4) ability to manage stressful situations, (5) performance in terms of safety, and (6) overall performance. Information was collected in these areas because previous research (Kantor et al., 1979) had shown these areas to be sensitive to instructor perception of sex differences. The surveys used in this study had been used successfully in prior research and were adapted with minor changes for this project. A more detailed description and an example of these surveys is provided in Kantor et al. (1979). Surveys were administered to the instructors immediately after the completion of that phase of training.

Statistical Procedure

To identify statistically significant differences between instructor ratings on men and women, _t_ ratios were computed and evaluated against critical _t_ values. For all _t_ tests, the Type I error rate (α) was controlled at .01 per family of comparisons using the Bonferroni technique (Miller, 1966). This procedure controls error rate at the determined level regardless of the number of comparisons or the degree of interdependency.

RESULTS AND DISCUSSION

Undergraduate Pilot Training (UPT)

As of July 1978, three groups of 10 women each had entered UPT. Survey results from the instructor pilots who taught members of the third group of women are presented in Table 1. Although the absolute values of the ratings for male co-students were somewhat better than for the females, statistical tests (also summarized in Table 1) were performed and none of the observed differences were found to be statistically significant. This is of particular interest because it was found that similar instructors of the first and second groups of women in UPT rated the male co-students significantly better on "ability to manage stressful situations" and "overall performance" (Kantor et al., 1979). The absence of significant differences among the instructor ratings from the third group of women to enter UPT might represent either a perceptive change on the part of the instructor corps or a difference in characteristics among the female student pilot groups. Additional research will continue to monitor the instructor ratings of men and women entering UPT, but overall, it would appear that little gender difference exists through this point in training.

Survival-Resistance Training

After completing UPT, all Air Force pilots attend Survival-Resistance training, a course designed to enhance a downed aircrew's capability to survive until rescued or maintain appropriate professional behavior in a prisoner of war situation. Survey results from the instructors who taught the female pilots, and their male co-students in Survival-Resistance training are presented in Table 2. Statistical testing revealed no significant gender difference among the instructor ratings of men and women completing this phase of training.

Replacement Training Units (RTU)

After completing Survival-Resistance training, Air Force pilots are sent to RTU where they are taught to fly the particular type of aircraft that they

Table 1

Instructor Ratings from UPT*

	Males		Females		
	$\bar{X}$	σ	$\bar{X}$	σ	t ratio[a]
1. Attitude towards instruction	1.5	.76	2.1	1.25	-1.16
2. Background knowledge pertinent to training	2.0	.76	2.4	.92	- .95
3. Ability tc acquire skills	2.3	.71	2.4	.75	.39
4. Ability to manage stressful situations	2.1	.35	2.9	.83	-2.51
5. Performance in terms of safety	2.0	.76	2.4	.92	- .95
6. Overall performance	2.0	.53	2.8	1.04	-1.94

Table 2

Survival Instructors Rating*

	Males		Females		
	$\bar{X}$	σ	$\bar{X}$	σ	t ratio[b]
1. Attitude towards instruction	2.36	1.0	1.91	.94	1.09
2. Background knowledge pertinent to training	3.00	.89	3.45	.69	-1.33
3. Ability to acquire skills	1.91	.70	1.91	.70	.00
4. Ability to manage stressful situations	2.55	.82	2.36	.81	.55
5. Performance in terms of safety	2.00	.63	1.82	.40	.80
6. Overall performance	2.18	.60	2.27	.81	-.30

Notes. 1. Survey instructions: Circle the number which best illustrates your rating on the following attributes for male and female pilots you have instructed.

2. Survey item options: Scale 1 (Extremely good) through 5 (Extremely poor).

[a] Bonferroni critical t = 4.64, α = .01, #c = 6, df = 7.

[b] Bonferroni critical t = 4.26, α = .01, #c = 6, df = 10.

*Lower scores reflect better ratings.

will be flying in their operational squadrons. Survey results from the RTU instructors are presented in Table 3. Statistical testing again revealed no significant differences between the instructor's ratings of male and female pilots. These comparisons were of particular interest since when the female pilots program began, there was some concern about how well a woman could handle some of the heavier Air Force aircraft. From the data available, it would appear that no serious problems were encountered and that the women assigned to heavy aircraft performed as well as males assigned to the same aircraft type.

In-Unit Performance

After completing RTU, Air Force pilots are sent to their operational squadrons, but advanced training in their unit aircraft is a continuing process. Instructor pilots in each operational squadron ensure that all pilots meet Air Force standards and continue to develop their pilot skills. Instuctor pilot ratings from the operational squadrons are presented in Table 4. Statistical testing revealed no significant differences between the female pilots and male fellow-pilots during their in-unit training. These comparisons are of interest because they represent a measure close to that of actual job performance including the ability of women to fly Air Force aircraft in actual operational environments. The finding of no gender differences in assigned unit performance appears to indicate that these women are capable of flying as operational Air Force pilots.

Combined Instructor Ratings

Since relatively small numbers of instructors were surveyed at any single point of training, it is possible that important differences might have been masked by the limited power of statistical testing with small _n_'s. Therefore, the survey results of all instructors combined were evaluated and are presented in Table 5. Statistical testing with this larger data base still revealed no significant differences among the instructor ratings of men and women in the Air Force pilot program.

CONCLUSION

In a prior phase of this project (Kantor et al., 1979), some significant differences were found between the instructor ratings of male and female student pilots in UPT. However, data presented in this report are notable because of the absence of any significant differences among instructor ratings of men and women Air Force pilots. No significant differences were found through UPT, Survival-Resistance training, RTU, or in-unit training. These similarities in ratings lead to the conclusion that men and women can be trained effectively using the existing UPT program and will perform similarly in advanced training and, at least initially, in their operational flying squadrons. Even considering the significant differences previously reported, the communalities between the sexes greatly outweigh any differences which have been found. Overall, it would appear that women constitute a viable and, as yet, still largely untapped aircrew resource for the Air

Table 3

Instructors Ratings from RTU*

	Males		Females		
	$\bar{X}$	σ	$\bar{X}$	σ	t ratio[a]
1. Attitude towards instruction	1.90	.56	1.70	.92	1.12
2. Background knowledge pertinent to training	2.18	.60	2.30	.90	- .69
3. Ability to acquire skills	2.02	.71	2.13	.77	- .66
4. Ability to manage stressful situations	2.16	.76	2.38	.85	-1.21
5. Performance in terms of safety	1.92	.80	1.86	.87	.32
6. Overall performance	2.24	.68	2.13	.78	.66

Table 4

In-Unit Instructors Ratings*

	Males		Females		
	$\bar{X}$	σ	$\bar{X}$	σ	t ratio[b]
1. Attitude towards instruction	2.00	.89	1.30	.81	1.43
2. Background knowledge pertinent to training	2.00	.89	1.80	.98	.37
3. Ability to acquire skills	1.30	.51	1.30	.51	.00
4. Ability to manage stressful situations	1.5	.83	1.6	1.03	- .19
5. Performance in terms of safety	1.8	.98	1.6	.81	.39
6. Overall performance	2.1	1.1	1.8	.75	.55

Notes. 1. Survey instructions: Circle the number which best illustrates your rating on the following attributes for male and female pilots you have instructed.

2. Survey item options: Scale 1 (Extremely good) through 5 (Extremely poor).

[a] Bonferroni critical t = 3.38, α = .01, #c = 6, df = 38.

[b] Bonferroni critical t = 6.14, α = .01, #c = 6, df = 5.

*Lower scores reflect better ratings.

Table 5

All Instructors Ratings*

	Males		Females		
	$\bar{X}$	σ	$\bar{X}$	σ	t ratio[a]
1. Attitude towards instruction	1.97	.73	1.75	.96	1.46
2. Background knowledge pertinent to training	2.28	.77	2.48	.98	-1.28
3. Ability to acquire skills	1.97	.71	2.05	.76	- .62
4. Ability to manage stressful situations	2.16	.77	2.37	.89	-1.43
5. Performance in terms of safety	1.94	.77	1.90	.81	.29
6. Overall performance	2.16	.65	2.22	.83	- .46

Notes. 1. Survey instructions: Circle the number which best illustrates your rating on the following attributes for male and female pilots you have instructed.

2. Survey item options: Scale 1 (Extremely good) through 5 (Extremely poor).

[a] Bonferroni critical t = 3.29, α = .01, #c = 6, df = 63.

*Lower scores reflect better ratings.

Force. Although additional research is needed to continue monitoring the performance of women Air Force pilots, at this point, the majority of indications are quite favorable for the continued expansion of the role of the female military pilot.

REFERENCES

Kantor, J.E., Noble, B.E., Leisey, S.A., & McFarlane, T. Air Force female pilots program: Initial performance and attitudes. AFHRL-TR-78-67. Brooks AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, Feb 1979.

Miller, R.G., Jr. Simultaneous statistical inference. New York: McGraw-Hill, 1966, 67-70.

MALE RECRUIT ATTITUDES TOWARD WOMEN
A FIVE-YEAR TREND STUDY*

Gerry L. Wilcove and Patricia J. Thomas

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

This study examined whether the attitudes of male recruits toward women have become more traditional or nontraditional over the past five years (1975-1979). It was expected that each successive yearly sample of recruits would evidence more liberal attitudes toward women than the previous sample because of the influence of the women's rights movement. This expectation was based on the concept of "cultural diffusion," which is the social evolution of a culture (in this case, our general American culture) in response to a counter-culture (i.e., the women's rights movement) (Weston, 1977).[1]

The study had two secondary goals: (1) to compare the attitudes of a male and female recruit sample in 1979, and (2) to compare the 1979 attitudes of a female recruit sample with the attitudes of a 1976 sample. It was expected that females would evidence less traditional attitudes in 1979 than males (cf. Savell & Woelfel, 1975)[2] and, consistent with cultural diffusion, female attitudes in 1979 were expected to be less traditional than female attitudes in 1976.

METHOD

Instrument

Item Content

Two types of items--demographical and orientational—were written for this study or taken from existing instruments; they were collectively termed the Sex Role Questionnaire (SRQ). Thirteen demographical items were included on race, hometown size, religion, etc.

Twenty-two orientational items were administered during all five years to men without any change in content, and an additional eight items were administered during the last four years. The orientational items, which use a Likert format, include belief statements (e.g., "Women are absent from work more often than men because of sickness") and, what might be termed, ideological statements (e.g., "A woman's place is in the home"). Both belief and ideological items

*Paper presented at the 21st Annual Conference of the Military Testing Association, San Diego, California, 15 October - 19 October 1979. The opinions and assertions contained herein are those of the senior author and are not to be construed as official or as reflecting the views of the Navy Department.

[1] Weston, L. The study of society. Guilford, Conn.: Bushkin, 1977.

[2] Savell, J. M. & Woelfel, J. C. Attitudes concerning job appropriateness for women in the Army. (ARI Research Memorandum 75-3). U. S. Army Research Institute for the Behavioral and Social Sciences, Arlington, VA, June 1975.

share a common trait: They are designed to reflect a traditional (T) or non-traditional (NT) attitude toward women. Tables 2 and 3 under "Results and Discussion" present the orientational items and whether they were classified into a T or NT category.

Validity of Classification System

Twelve judges were asked to classify the items as T, NT, or "neither." Ten or more of these judges classified 20 of the 30 items as T or NT, and two-thirds classified 28 of 30 as T or NT. (The two items failing the "two-thirds" criterion were dropped from the study.) Despite this support, the judges and I (the senior author) may define "traditional" and "non-traditional" differently and thus might classify the items differently. For me, items expressing gender-specific (person-specific) ideas on roles, competencies, traits, and rewards are T (NT). I accordingly classified some items as T and others as NT. On the average, the judges and I agreed on 22.8 of 28 items. It was concluded that sufficient evidence exists for using the T/NT classification system in the study.

Prediction of Behavior from Attitudes

The present study simply measured attitudes; however, it is believed that results have implications for behavior. In particular, consider Ajzen and Fishbein's (1977) discussion of prejudice after an exhaustive review of the attitude-behavior literature:[3]

> If the investigator is really interested in a general behavioral pattern, such as discrimination toward blacks, the behavioral criterion should involve observation of different discriminatory behaviors toward various black individuals in a variety of contexts. A general measure of attitudes toward blacks will correspond to such a criterion. (p. 913)

The SRQ is a general measure of attitudes toward women, primarily the attitudes of men. Based on the conclusion of Ajzen and Fishbein, it is reasonable to expect that a male with a traditional attitude will exhibit a negative pattern of behavior toward females in nontraditional roles, and a male with a nontraditional attitude will exhibit a positive behavioral pattern.

Samples and Data Collection

The SRQ was administered to male recruits at RTC San Diego in August 1975, August 1976, March and April 1977, August through November 1978, and July and August 1979. (The negative consequences, if any, of not administering the SRQ at exactly the same time of year each time are discussed under "Results and Discussion.") Usable questionnaires were obtained from 801 (1975), 1090 (1976), 929 (1977), 846 (1978), and 607 (1979) individuals, respectively. Usable data were also obtained in August 1979 from 553 male recruits and 234 female recruits in Orlando, FL and from female recruits (N = 450) in Orlando in June and July 1976.

[3] Ajzen, I., & Fishbein, M. Attitude-behavior relations: A theoretical analysis and review of empirical research. Psychological Bulletin, 1977, 84, 888-918.

Data Analysis

Molecular Approach

Item means and standard deviations were computed for each sample. The *t* test for independent samples was conducted between the 1979 Orlando male recruit and female recruit samples. It was also conducted between the 1979 and 1976 female recruit samples. (The trend analysis is described next under "Molar Approach.")

Molar Approach

After examining item means, the *t* test for a curvilinear trend with unequal *N*'s was used to analyze the data from the five yearly male samples (Edwards, 1968 pp. 136-137).[4] Responses to an attitude scale served as the data in this analysis. The 20 items which had been administered all five years and which had passed the previously described validity tests comprised the scale.[5] The equation for computing the *t* statistic is as follows:

$$t = \frac{k(2\bar{X}_1 - \bar{X}_2 - 2\bar{X}_3 - \bar{X}_4 + 2\bar{X}_5)}{\sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2 + (n_5 - 1)s_5^2}{n_1 + n_2 + n_3 + n_4 + n_5 - 5}\right]\left[\frac{4}{n_1} + \frac{1}{n_2} + \frac{4}{n_3} + \frac{1}{n_4} + \frac{4}{n_5}\right]}}$$

Where

k = number of items in attitude scale

$\bar{X}_i$ = the average of all item means in the attitude scale for a given yearly sample

[4] Edwards, A. L. Experimental design in psychological research. New York: Holt, Rinehart, and Winston, 1968.

[5] "Orthogonal" factor analyses were conducted in 1976 of male and female responses, and uninterpretable factors were obtained. These negative findings were expected since the questionnaire was designed to measure a single traditional/nontraditional dimension. No additional factor analyses were conducted (e.g., oblique) because even if subscales were obtained for a given year, it is extremely unlikely that these scales would contain the same exact items across all five years.

n_1 = number of male recruits in 1975 sample

n_2 = number in 1976 sample

n_3 = number in 1977 sample

n_4 = number in 1978 sample

n_5 = number of 1979 sample

s_i^2 = the variance of the scale scores for a given sample, where a scale score was obtained by coding response options from 1 to 5 for the T items, reverse scoring the NT items, and adding the coded responses

RESULTS AND DISCUSSION

Table 1 presents the yearly percentage of men agreeing (A) or disagreeing (D) with each orientational item[6]; the eight items which were not administered in 1975 are deleted. For the traditional items, the "strongly disagree" and "disagree" responses are combined; for the non-traditional or "contemporary" items, the number of "strongly agree" and "agree" responses are combined. The table's right-hand column indicates whether a curvilinear or linear trend seems to exist; or, if neither situation prevails, "no trend" is indicated.

Eleven of the 20 items yielded percentages which approximate a curvilinear trend; for example, the orientation was initially traditional (1975), became more traditional (1976 and 1977), and then less traditional (1978 and 1979). The grand mean of the item means for the attitude scale was 3.40 for 1975, 3.26 for 1976, 3.29 for 1977, 3.37 for 1978, and 3.35 for 1979. The higher the mean, the more nontraditional the attitude. The t test conducted for the means yielded a value of 4.359 ($p < .0005$, df=4268), thus indicating a significant curvilinear trend.

Is this significant result an artifact of the large N's and thus invalid? No! A large N permits the detection of a small but stable trend, but does not create one where none exists. Are results statistically significant but meaningless in a practical sense? I do not believe so: One cannot assume that yearly differences in orientations must be of a certain magnitude to infer differences in behavior. Furthermore, obtained yearly percentage differences compare favorably with other trend studies (cf. Crawford, Thomas, & Thomas, 1976)[7]

[6] Percentages are given instead of means, although means are used in the Edward's trend equation--the differences between percentages are easier to interpret than the differences between means (e.g., 3.52 vs 3.38). Since the response scale had a neutral midpoint, percent agreeing is not the reciprocal of percent disagreeing.

[7] Crawford, K. S., Thomas, P. J., & Thomas, E. D. Preservice drug usage among naval recruits: A 5-year trend analysis. (NPRDC TR 76TQ-45) Navy Personnel Research and Development Center: San Diego, September 1976.

Table 1

Yearly Percentage of Male Recruits Disagreeing (D) with Traditional (T) Items and Agreeing (A) with Non-Traditional (NT) Items

Item Number/Content		1975	1976	1977	1978	1979	Apparent Trend[a]
		Traditional Items					
1T.	Women aren't serious about jobs (D)	70.2	61.8	60.1	66.3	70.3	CU
2T.	Women too uncoordinated to do men's job (D)	67.6	57.1	55.0	60.4	55.2	No trend
3T.	Would not want to fly in plane piloted by woman (D)	64.0	56.7	55.9	65.5	58.7	No trend
4T.	Cost of training women wasted (D)	63.3	52.3	49.8	46.5	48.0	LT
5T.	Women's job mistakes more excusable than men's (D)*	63.3	55.8	57.5	59.9	59.8	CU*
6T.	Married women shouldn't work (D)	62.3	45.6	47.4	52.9	55.1	CU
7T.	Women should stick to "women's jobs" (D)	51.5	39.3	38.6	47.7	47.2	CU
8T.	Tough competitive jobs not for women (D)	47.2	33.7	39.2	43.5	43.1	CU
9T.	Harder for women to make decisions than men (D)	43.5	31.6	37.2	43.7	39.5	No trend
10T.	Woman's place is in home (D)	40.0	40.9	36.9	44.1	39.2	No trend
11T.	Women out sick more than men (D)	36.5	27.8	30.1	35.8	34.4	CU
12T.	Qualified man should get job over qualified woman (D)	30.7	30.8	33.0	29.5	39.7	LC
13T.	Children of working mothers feel unloved (D)	28.7	21.8	19.4	22.7	25.5	CU
14T.	Women are weaker sex (D)	24.1	21.5	23.5	27.3	28.9	CU
15T.	Is time of month when women's emotions interfere with job (D)	17.0	10.9	10.0	12.9	15.8	CU
16T.	Women more emotional than men (D)	13.4	9.9	10.1	14.4	13.6	CU
		Non-Traditional Items[b]					
1NT.	Girls need as much education as boys (A)	83.2	79.6	80.8	81.5	82.7	CU
2NT.	Unfair to promote on basis of sex (A)	82.2	71.1	69.7	73.9	72.5	CU
3NT.	Women do jobs as well as men (A)	73.3	74.5	82.9	79.3	80.2	LC
4NT.	If mother can manage family and job, should be allowed to work (A)	71.4	72.7	78.5	76.3	74.8	No trend
5NT.	Women raise morale of workgroup (A)*	61.2	49.7	56.0	55.5	51.7	No trend
6NT.	No reason father shouldn't care for children (A)	45.9	49.5	53.4	55.7	54.5	LC

Note. All data were collected at RTC San Diego. Since the response scale had a neutral midpoint, percentage disagreeing is not the reciprocal of percentage agreeing.

[a] CU = curvilinear trend, LT = linear trend toward a more traditional orientation, LC = linear trend toward a more non-traditional or "contemporary" (C) orientation.

[b] The table includes many more traditional items than non-traditional ones, suggesting the possibility of a response set. When one considers, however, the eight items that were deleted from the analyses (i.e., those not administered in 1975), the breakdown becomes 14 non-traditional items and 16 traditional items.

*Did not meet the "two-thirds" validity criterion established for the traditional/non-traditional classification system and thus was not included in the trend analysis of scale scores. It is included, however, in the table for interested readers.

Since survey data were not collected at the same exact time of year each time, should demographical differences among samples be partialled out? After all, such differences may signify that samples from different populations are being compared (i.e., "apples are being compared with oranges")! I do not believe so, because recruit populations do in fact vary yearly: (1) on their knowledge of word definitions, and (2) for NTC San Diego recruits, on race and aptitude. In addition, for any obtained trend to reflect sample differences in demographics, the demographics would need to be distributed curvilinearly across yearly samples and to be correlated appreciably with attitude. However, 12 of the 13 demographic variables were not distributed curvilinearly, and the 13th typically correlated .03 with the attitude items.

The shift toward more traditional attitudes, which was detected between 1975 and both 1976 and 1977, was in part replicated by the General Social Survey.[8] In this survey, a representative national sample responded to 4 items on women's roles; more traditional attitudes were found in 1977 than in 1975 on two of the items. (No change was evidenced on the other two items.)

Tables 2 and 3 present the results for the 1976 and 1979 female recruit samples. For 12 of the 28 items, women exhibited a significantly more traditional attitude on sex roles in 1979 than in 1976. Tables 4 and 5 compare the attitudes of the 1979 Orlando male and female recruit samples. In every instance, females were significantly less traditional than males.

CONCLUSIONS AND RECOMMENDATIONS

1. It is tentatively concluded that the attitudes of male recruits towards women are not becoming more liberal with each passing year, and it would be a mistake on the part of the Navy to assume so.

2. While it appears that female recruits are becoming more traditional in their attitudes towards women's roles, they are still markedly more liberal than males.

3. The results of this study should be distributed to:

a. The Women at Sea Workshops which are conducted by HRM Specialists solely for women. One of the primary purposes of these workshops is to discuss the fears and expectations that women have with respect to working with men.

b. The Women in the Navy Workshops which are conducted by the HRM Specialists for middle managers who have received, or are about to receive, women onboard. One of the primary purposes of these workshops is to create an awareness of issues pertinent to women in the Navy.

c. The Defense Race Relations Workshops conducted for EEO Specialists. These workshops have limited information available to them on factors affecting integration of women into the Navy.

[8] Inter-University Consortium for Political and Social Science Research, Institute for Social Research, University of Michigan, Ann Arbor.

Table 2

Women's Sex Role Perceptions: Percentages Disagreeing with Traditional (T) Items in 1976 and 1979

Item Number/Content		Percentage Disagreeing		Mean[1]	S. D.	Mean	S. D.	t Value
		1976	1979	1976		1979		
4T.	Cost of training women wasted	89.8	68.0	4.40	.85	3.86	1.13	6.39[d]
11T.	Women out sick more than men	70.0	47.4	3.92	1.06	3.38	1.11	6.11[d]
5T.	Women's job mistakes more excusable than men's*	90.0	77.5	4.34	.75	4.04	.98	4.05[d]*
15T.	Is time of month when women's emotions interfere with job	58.4	45.2	3.59	1.15	3.25	1.18	3.59[d]
6T.	Married women shouldn't work	89.8	83.7	4:39	.80	4.16	.96	3.13[c]
16T.	Women more emotional than men	30.4	17.3	2.80	1.22	2.55	1.10	2.70[c]
9T.	Harder for women to make decisions than men	72.2	62.0	3.95	1.05	3.72	1.08	2.65[c]
7T.	Women should stick to "women's jobs"	85.7	78.5	4.28	.92	4.08	1.07	2.42[b]
10T.	Woman's place is in home	54.0	60.9	3.65	1.04	3.84	1.01	2.30[a]
2T.	Women too uncoordinated to do men's jobs	91.7	86.3	4.43	.83	4.28	.85	2.18[a]
13T.	Children of working mothers feel unloved	53.5	48.3	3.55	1.16	3.35	1.22	2.06[a]
3T.	Would not want to fly in plane piloted by woman	79.9	76.8	4.17	.99	4.06	.94	1.42
1T.	Women aren't serious about jobs	90.7	88.8	4.50	.92	4.41	.98	1.16
8T.	Tough competitive jobs not for women	81.3	77.3	4.11	.91	4.03	.90	1.09
12T.	Qualified man should get job over qualified woman	79.3	77.8	4.27	.90	4.19	.94	1.07
14T.	Women are weaker sex	-	60.0	-	-	3.71	1.09	-

Note. *N* for 1976 (without subtracting missing responses) = 450; comparable 1979 *N* = 234. There were typically 1 or 2 missing responses; at most 16. T values were based on *N*'s which excluded the missing responses. All data were collected at RTC Orlando.

[1] Coding: Scale ranged from 1 = strongly agree, 3 = neither agree nor disagree, 5 = strongly disagree

[a] *p*<.025

[b] *p*<.01

[c] *p*<.005

[d] *p*<.0005

All tests were one-tailed.

*Did not meet "two-thirds" validity criterion established for classification system.

Table 3

Women's Sex Role Perceptions: Percentages Agreeing
With Non-Traditional (NT) Items in 1976 and 1979

Item Number/Content		Percentage Agreeing		Mean[1]	S. D.	Mean	S. D.	t Value
		1976	1979	1976		1979		
13NT.	Women as logical as men	90.2	88.6	1.69	.83	1.81	.75	1.89[a]
5NT.	Women raise morale of workgroup*	52.3	52.9	2.37	.80	2.49	.86	1.75[a]*
11NT.	Men influenced as easily as women	75.1	80.2	2.02	.98	1.89	.92	1.70[a]
1NT.	Girls need as much education as boys	94.5	91.8	1.38	.69	1.47	.76	1.51
4NT.	If mother can manage job and family, should be allowed to work	93.1	90.1	1.65	.78	1.70	.79	.78
6NT.	No reason father shouldn't care for children	82.0	75.6	1.92	.98	2.04	.97	1.52
7NT.	Women should play contact sports	46.2	40.5	2.66	1.03	2.75	1.04	1.07
8NT.	Female mechanics as good as male mechanics	83.7	82.7	1.84	.93	1.82	.85	.28
9NT.	O.K. for woman to drive truck or bus	92.2	94.0	1.72	.74	1.69	.73	.50
10NT.	Women should have same privileges/ responsibilities as men	68.6	64.7	2.15	.99	2.25	1.15	1.12
2NT.	Unfair to promote on basis of sex	91.3	88.4	1.58	.94	1.55	1.04	.36
12NT.	Women carry own weight in workgroup	83.6	83.2	1.87	.84	1.90	.78	.46
14NT.	Men panic in crises as often as women	85.1	83.3	1.78	.93	1.85	.88	.95
3NT.	Women do jobs as well as men	-	88.9	-	-	1.72	.77	-

Note. Base level N for 1976 = 450; for 1979 = 234. All data were collected at RTC Orlando.

[1] Coding: Scale ranged from 1 = strongly agree, 3 = neither agree or disagree, 5 = strongly disagree

[a] $p < .025$, one-tailed test

*Did not meet "two-thirds" validity criterion established for classification system.

Table 4

1979 Sex Role Perceptions: Percentage of Males (M) vs Females (F) Disagreeing with Traditional (T) Items

Item Number/Content	Percentage Disagreeing		Mean[1]	S. D.	Mean	S. D.	t Value
	M	F	M		F		
10T. Woman's place is in home	33.5	60.9	3.10	.97	3.84	1.01	9.47[d]
12T. Qualified man should get job over qualified woman	36.1	77.8	3.20	1.01	4.19	.94	11.85[d]
15T. Is time of month when women's emotions interfere with job	13.6	45.2	2.47	.96	3.25	1.18	8.89[d]
1T. Women aren't serious about jobs	70.5	88.8	3.77	1.02	4.41	.98	8.24[d]
5T. Women's job mistakes more excusable than men's*	56.8	77.5	3.51	1.18	4.04	.98	6.42[d]
11T. Women out sick more than men	25.4	47.4	2.97	.88	3.38	1.11	5.00[d]
6T. Married women shouldn't work	50.2	83.7	3.35	1.03	4.16	.96	10.55[d]
16T. Women more emotional than men	9.4	17.3	2.10	.99	2.55	1.10	5.36[d]
8T. Tough competitive jobs not for women	42.8	77.3	3.17	1.09	4.03	.90	11.45[d]
7T. Women should stick to "women's jobs"	40.6	78.5	3.13	1.05	4.08	1.07	11.42[d]
13T. Children of working mothers feel unloved	24.6	48.3	2.78	1.07	3.35	1.22	6.18[d]
9T. Harder for women to make decisions than men	36.5	62.0	3.06	1.05	3.72	1.08	7.86[d]
3T. Would not want to fly in plane piloted by woman	54.7	76.8	3.44	1.08	4.06	.94	8.06[d]
14T. Women are weaker sex	24.4	60.0	2.76	1.08	3.71	1.09	11.14[d]
4T. Cost of training women wasted	47.8	68.0	3.31	1.03	3.80	1.13	6.36[d]
2T. Women too uncoordinated to do men's jobs	55.1	86.3	3.43	1.02	4.28	.85	11.87[d]

Note. Base level N for males was 553 and for females 234. T values were based on N's which excluded missing responses, at most 11 for males and 18 for females. All data were collected at RTC Orlando.

[1] Scale ranged from 1 = strongly agree, 3 = neither agree nor disagree, 5 = strongly disagree

[a] $p < .025$

[b] $p < .01$

[c] $p < .005$

[d] $p < .0005$

All tests were one-tailed.

*Did not meet validity criterion established for classification system.

Table 5

1979 Sex Role Perceptions: Percentage of Males (M) vs Females (F) Agreeing with Non-Traditional (NT) Items

Item Number/Content	Percentage Agreeing		Mean[1]	S. D.	Mean	S. D.	t Value
	M	F	M		F		
1NT. Girls need as much education as boys	84.8	91.8	1.82	.91	1.47	.76	5.53[d]
4NT. If mother can manage job and family, should be allowed to work	77.5	90.1	2.06	.97	1.70	.79	5.43[d]
6NT. No reason father shouldn't care for children	51.9	75.6	2.69	1.23	2.04	.97	7.89[d]
7NT. Women should play contact sports	30.3	40.5	3.28	1.29	2.75	1.04	6.04[d]
8NT. Female mechanics as good as male mechanics	57.0	82.7	2.54	1.08	1.82	.85	9.92[d]
9NT. O.K. for woman to drive truck or bus	81.1	94.0	2.06	.81	1.69	.73	6.26[d]
10NT. Women should have same privileges/ responsibilities as men	56.0	64.7	2.52	1.10	2.25	1.15	3.03[c]
2NT. Unfair to promote on basis of sex	78.9	88.4	1.91	1.00	1.55	1.04	4.47[d]
11NT. Men influenced as easily as women	57.7	80.2	2.48	1.03	1.89	.92	7.88[d]
12NT. Women carry own weight in workgroup	62.4	83.2	2.38	.85	1.90	.78	7.64[d]
5NT. Women raise morale of workgroup*	50.7	52.9	2.49	.86	2.37	.80	1.85*
13NT. Women as logical as men	61.2	88.6	2.46	.92	1.81	.75	10.20[d]
14NT. Men panic in crises as often as women	58.4	83.3	2.56	1.09	1.85	.88	9.46[d]
3NT. Women do jobs as well as men	71.6	88.9	2.14	.95	1.72	.77	6.30[d]

Note. Base level *N* for males was 553 and for females, 234. T values were based on *N*'s which excluded missing responses. All data were collected at RTC Orlando.

[1] Scale ranged from 1 = strongly agree, 3 = neither agree nor disagree, 5 = strongly disagree

[a] $p < .025$

[b] $p < .01$

[c] $p < .005$

[d] $p < .0005$

All tests were one-tailed.

*Did not meet validity criterion established for classification system.

OFFICER TASK ANALYSIS I

Panel Discussion: Adaptive Approaches to Officer Task Analysis: Planning Flexibility for Diverse Applications

Chairperson: Dr. John B. Mocharnuk, McDonnel Douglas Astronautics Company

Participants:

Judy Akin, U.S. Army Training and Doctrine Command

H. W. Ruck, Air Force Human Resources Laboratory

B. M. Berger and D. Worstine, U. S. Army Military Personnel Center

S. J. Van Nostrand and R. O. Waldkoetter, U. S. Army Research Institute

R. A. Marco and J. B. Mocharnuk, McDonnel Douglas Astronautics Company.

The discussants will examine issues related to officer task analysis, focusing on innovative techniques which have emerged during the course of recent officer task analysis research and development. As procedures for conducting officer job analysis have developed attempts to standardize these procedures have, in some cases, led to a rigidity which allows one to ignore important facets of a job or, alternatively, cause a level of task specificity which exceeds that required by users of the task analysis data. When undue rigidity is encountered it may be accompanied by a belief that the task analysis is an end product rather than a means to achieving some other end such as proper training content, selection test items, or improved equipment design. The panel will discuss issues related to the design of task analysis procedures and data sets which are appropriate for diverse jobs and can be adapted for special applications. Insights emerging from a developmental officer analysis program and implications from a recent analysis which included task difficulty by scenario and task criticality by scenario ratings will be presented.

PUBLIC SECTOR EXECUTIVE JOB CONTENT AND TRAINING NEEDS RELATIONSHIP

Laurie Broedling
Alan Lau

Navy Personnel Research and Development Center

Arthur Newman
Paula Harvey

San Diego State University

INTRODUCTION

This paper is part of a symposium pertaining to officer job analysis. At present, there is no job analysis being conducted in the Navy on officers' jobs; the job analysis program has been restricted to enlisted personnel. We have, however, recently conducted an extensive study of the Navy career civilian executive job, i.e., the Navy's supergrades.[1] The majority of this study was devoted to performing a job analysis of these positions. The results of that analysis will be described in this paper. The relevance of these findings to this particular symposium is twofold. First, there is commonality of many functions and tasks between senior career civilians and military officers. Second, due to the interlocked nature of many senior civilian and military positions, any comprehensive job analysis of one set requires at least an informal or indirect analysis of the other set. Consequently, in the process of doing our study, we did acquire some information about the job functions of senior military officers.

Job analysis varies on a number of different dimensions, including the purposes for which it is conducted, the conceptual approaches available, and the ways to collect the data. With respect to the purposes, job analysis information can be used to help develop personnel selection systems, performance appraisal systems, training systems, man-machine systems, as well as for a variety of other specific management decisions, such as the redesign of work flow. In the military, the vast bulk of job analysis which is conducted has as its primary purpose the development of training systems. In other words, the job analysis information is used as the basis for curriculum development. Our particular study also had the identification of training needs as its primary purpose. The basic presumption was that it is difficult to design useful executive development and training programs without knowledge of what the executive job entails. Despite the seeming obviousness of that presumption, the fact is that training programs for managerial types of jobs in both the private and public sector have very rarely been based on job analysis. Most management training has been based on armchair speculation regarding what managers do or should do. Job analysis has traditionally been applied to jobs that are primarily nonmanagerial in nature. It is one of the contentions of this paper that the

[1] The full study is described in summary form (Broedling & Lau, 1979) and in a full technical report (Lau, Broedling, Parisi, Newman, & Harvey, 1979).

approaches and data collection methods selected to analyze managerial jobs should perhaps be considerably different than those used for nonmanagerial jobs.

There are a number of other characteristics of the voluminous body of literature on management which meant most of it could not be directly applied to our research problem. First, most of it has developed with the private sector manager in mind. Executive behavior has received considerably less systematic attention in the public sector. Second, little of it pertains specifically to top executives; either it pertains to middle or first-level supervisors, or it treats management as a function that is the same across all hierarchical levels, tasks, or technologies. Third, most of the empirical research has dealt with only one aspect of management, namely, leadership. Although leadership does represent a major managerial function, it is by no means the only one. Fourth, management and leadership theories have been short-range and atomistic, focusing on leader-group rather than leader-group system relationships (McCall & Lombardo, 1978) thus slighting the importance of the organizational environment.

There are a variety of different approaches for doing job analysis (McCormick, 1976), such as the functional approach, the worker-oriented approach, the abilities approach, and the task approach. Most military job analysis has employed the task approach, in which the job is broken down into its component tasks. Since no job analysis work has ever been done in senior military management jobs and little has been done in management jobs in general, in our study we were deliberately eclectic, using some elements of all the available approaches.

There is also a variety of methods available to collect job analysis data (McCormick, 1976). Each method has different advantages and disadvantages. Despite the variety of methods available, most job analysis has been conducted using structured surveys. This holds particularly true for military job analysis. Moreover, the large majority of all types of management and leadership research has been done using structured surveys. This methodological restriction has been identified as a severe shortcoming of the leadership literature, and critics have called for the need for multimethod studies. One particularly articulate critic of the over-reliance on the survey has been Mintzberg (1973). He conducted a job analysis of executives by individually observing five different executives for one week each. Out of those observations, he developed ten categories of managerial job activities. Most importantly, his findings seriously challenged a number of "truisms" about management which had evolved based on survey findings and armchair speculation. In essence, Mintzberg challenged the notion that managers do long-range planning, perform careful problem analysis, do systematic resource allocation, etc. Instead, he found managers operate in a short-fused, crisis management environment which requires quick reactions based on minimal information. The result is they do little planning, communicate primarily orally, have little control over how they spend their time, and have a daily work routine which is very fragmented, hectic and diverse. Therefore, in our study we were eclectic in our data collection methods, deliberately invoking a multimethod design.

As mentioned above, the most frequent purpose for undertaking job analysis is to determine training needs. However, usually this means determining only training content, with little or no attention paid to determining training methods.

Part of the reason for this is that much job analysis is focused on job content, i.e., what functions the incumbents perform, and job characteristics are ignored, i.e., how the incumbents perform those functions. Yet the "how" might be equally as important as the "what." In turn, knowing about the job characteristics may help provide information about which training methods are most desirable. Particularly in the management area, determining training methods is an important aspect. Certain management skills, such as counseling, are probably better trained using experiential methods than formal, lecture methods. Therefore, it is important that managerial job analysis be designed to generate information about training methods as well as training content.

Another characteristic of most job analysis is that it is focused directly on the job activities or on the skills needed to perform those activities. It does not ordinarily include any direct assessment of the job environment. Yet the job environment frequently has a profound impact on people's performance of their job functions. Therefore, a description of the job environment can aid greatly to the analytical aspects of the job analysis. In fact, Mintzberg's (1973) findings indicate that the knowledge of the job environment is an indispensable part of describing and understanding the managerial job. In our study, we did spend a lot of time learning about the job environment of Navy civilian executives. While we began our first data collection efforts by focusing on job activities themselves, we rapidly found that it was impossible to adequately characterize these activities without learning about the larger environment in which they take place. For example, the fact that the Navy is in an environment of shrinking resources and the fact it is increasingly centralized both have a profound impact on its executives' jobs.

APPROACH

Subjects

The study was conducted on an executive population in the federal government—the highest graded civilians working for the U.S. Navy who are GS-16, 17, 18 or equivalent Public Law positions (N = 370). In addition, information was gathered from those military executives in the Navy's shore establishment who are the superiors of civilian executives. The purpose of the latter was to gather information about the working relationships among the two groups and to obtain military executives' perceptions of the job content, job characteristics, training needs, etc., of civilian executives.

Data Collection Methods

A multimethod approach was used in this study. The methods included interviews, observation, work activity diaries, and structured questionnaires. The first three methods were used on a subsample of the population, while the questionnaires were sent to all Navy career civilian executives. Data were collected between July 1977 and March 1978.

Semistructured interviews were conducted with 57 career civilian executives. These interviews averaged 1-1/4 hours. The major topics included executive job activities, training needs and experiences, selection, job environment, appraisal, and the civilian-military interface. Shorter interviews were conducted with 17 military executives.

A cross-section of 19 civilian executives were asked to describe their activities over a two-week period using a work diary form. A work sampling technique was used so that executives recorded activities performed on alternate four-hour morning or afternoon time blocks.

Four Navy civilian executives were observed doing their job over a two-day period. The observer used the Executive Work Diary Form to record every activity in which the executive engaged. The observer was as unobtrusive as possible, conversing with the executive only when information was needed about a given activity.

The Executive Questionnaire was developed using information gained from interviews and from the management literature. Using Mintzberg's framework of managerial activities (Mintzberg, 1973), 50 items describing work content were asked (1) how much time they spent on the average in each activity, and (2) how important each activity was to them in the successful conduct of their work. Responses were made on an 8-point scale, where 0 = "None" and 7 = "Great Deal." There were also 15 items assessing perceived job characteristics, and a variety of items assessing job environment.

The population of 370 Navy civilian executives was mailed a questionnaire. A total of 210 questionnaires were completed and returned. Those who returned the questionnaires were representative of the full population in terms of GS level, occupational series, and organization affiliation.

A shorter questionnaire was mailed to all military superiors of civilian executives (N = 98). Seventy percent of this population completed the questionnaire.

RESULTS

The first major finding was a tremendous amount of commonality across executive jobs in all aspects studied. For example, the questionnaire results were analyzed to see if responses differed along five different dimensions: (1) line vs. staff jobs; (2) science/engineering management vs. other types of management; (3) headquarters vs. field location; (4) job title; (5) type of supervisor. Very few statistically significant differences were found beyond the number otherwise expected by chance.

Job Content

It was found that Navy civilian executives perform all the types of tasks one ordinarily associates with the concept of "executive," such as resource allocation, leadership, information gathering and dissemination. They do not generally function just as advisors or staff members to senior military executives; they have considerable authority in their own right. The only major distinguishing feature in the job content of Navy civilian executives is that they function within the technical or functional specialty which has been their career field (e.g., physics, engineering, personnel).

Table 1 provides the 50 items describing work content in order of their mean rated importance. These data are presented in this paper because they most closely resemble the type of data collected in a conventional military task analysis. These results have some interesting properties. For example, the correlations between time spent and importance for each item vary tremendously, from .10 to .70. Also, all items have higher mean importance ratings than time spent ratings.

These items were also factor analyzed to investigate the basic dimensions associated with executive jobs in terms of importance. A principal component solution was obtained. Orthogonal rotation to the varimax criteria yield four major factors which accounted for 76 percent of the variance.[2] Since the same analysis done on the time dimension loaded on the same factors and was consistent, only results from the importance dimension will be described. The items that loaded on these factors will be described below. Thirty-seven out of the 50 items loaded at least on one of the factors.

Leadership, which entails staffing; the guidance, motivation, and development of subordinates; programming work; and resolving conflicts.

Executive decision-making, which entails policy-making, implementing directives, evaluating outcomes, and planning.

Technical problem solving, which entails directing, conducting, consulting, and reviewing the technical aspects of one's area of work specialization.

Information seeking and disseminatio, which entails receiving and transmitting information between the executive's organizational unit and the outside world.

Job Characteristics

One major characteristic is job variety, with few limits on what executives are required to do, except in terms of their working within their technical or functional specialty. Another characteristic is that executives spend the large majority of time interacting with other individuals. An analysis of the work diaries, for instance, indicated only 20% of their time was spent alone. A third characteristic is the long working hours, with executives reporting an average of 52 hours per week at the office and an additional eight at home. A fourth characteristic was the fragmented and hectic nature of the job, with the executive moving rapidly from one activity to another. Further, the large majority of executives report they do not exercise sufficient control over their own time. Executives jobs are also pressured and stressful. Almost 90% of the civilian questionnaire respondents reported that there is either moderate or great pressure on them to produce; none said that there was no stress on their jobs. Lastly, a fair amount of job sharing was identified, with 60% of the civilian executives saying they share their job responsibilities with one or more people, excluding their officially designaged department or division heads. Much of this job sharing is done with senior military officers.

[2] Since the number of questionnaire items relative to sample size is rather large, one may question the stability of these factors. Veldman's (1967) RELATE program was used to compare the underlying structural properties of executive job content. The entire sample (N = 210) was split into 2 random samples and separate factor analyses were performed. The two factor structures were then compared. Correlations between the factor variables derived from the two analyses ranged from .89 to .99, indicating a stable factor structure across both groups.

Table 1

Ranked Order of Job Activities Performed by Executives

Role	Item	Importance		Time		Correlation between ti and importance
		Mean[a]	S.D.	Mean[a]	S.D.	
Resource Allocator	Determining the long-range plans and priorities of your unit.	5.8	1.3	3.2	1.8	.55*
Leader	Evaluating the quality of subordinate job performance and providing recognition, encouragement, or criticism.	5.8	1.5	2.9	1.7	.32*
Leader	Providing guidance and direction to your subordinates.	5.8	1.6	4.2[b]	1.8	.55*
Resource Allocator	Allocating resources (manpower, money, material) among programs or units.	5.7	1.6	3.7[b]	2.0	.49*
Resource Allocator	Allocating your own time.	5.7	1.8	1.7	1.5	.19*
Disseminator	Keeping members of your unit informed of relevant information through meetings, conversations, and dissemination of written information.	5.6	1.5	3.3[b]	1.6	.36*
Monitor	Learning about fleet requirements and needs.	5.6	1.8	2.7	1.7	.44*
Leader	Attending to staffing requirements in your unit such as hiring, firing, promoting, and recruiting.	5.4	1.6	2.7	1.7	.22*
Resource Allocator	Participating in defining command strategies and policies.	5.4	1.7	2.7	1.6	.43*
Technical Expert	Judging the accuracy of approach and utility of technical programs and proposals.	5.3	1.8	3.1	2.0	.57*
Leader	Keeping abreast of who is doing what in your unit or command.	5.2	1.6	3.6[b]	1.7	.52*
Disturbance Handler	Taking immediate action in response to a crisis or "fire drill."	5.1	1.9	3.8[b]	2.0	.10
Technical Expert	Providing technical quality control through the review process.	5.1	2.0	3.1	2.0	.55*
Spokesman	Keeping sponsors, consumers, or other important governmental groups informed about your unit's activities and capabilities.	4.9	1.7	3.1	1.8	.52*
Monitor	Staying tuned to what is going on in outside organizations, including the professional and scientific communities.	4.9	1.7	2.8	1.7	.47*
Disturbance Handler	Resolving conflicts either within your unit or between your unit and other organizational components.	4.9	1.9	2.4	1.6	.28*
Liaison	Developing personal relationships with people outside your unit who sponsor your work or services.	4.8	2.0	2.7	1.4	.59*
Disturbance Handler	Preventing the loss or threat of loss of resources valued by your unit.	4.8	2.3	2.3	1.7	.49*
Leader	Attending to the training and development needs of your employees.	4.7	1.8	2.3	1.3	.48*
Spokesman	Defending your unit's projects and activities to other groups.	4.7	2.1	2.8	1.8	.43*
Resource Allocator	Programming work for your unit (what is to be done, when, and how) and assigning people to work on it.	4.7	2.1	2.4	1.7	.51*
Technical Expert	Consulting with others on technical matters.	4.6	1.8	2.9	1.7	.61*
Entrepreneur	Exploiting or initiating opportunities to improve or expand as a unit.	4.4	2.0	2.5	1.8	.57*
Disseminator	Implementing the directives of higher authorities.	4.4	2.1	3.0	1.9	.31*

[a] Based on an 8-point scale, where 0 = None and 7 = Great Deal.

[b] One of five highest ranked items as to time spent.

*$p < .05$.

Table 1 (Continued)

Role	Item	Importance		Time		Correlation between tin and importance
		Mean[a]	S.D.	Mean[a]	S.D.	
Negotiator	Negotiating with groups internal to your command for necessary materials, support, commitments, etc.	4.4	2.0	2.5	1.8	.48*
Leader	Integrating subordinates' goals (e.g., individual development plans, career goals, work preferences) with the command's work requirements.	4.4	2.1	1.8	1.4	.59*
Disseminator	Transmitting ideas and information from your outside contacts to appropriate people inside your command.	4.4	1.9	2.1	1.4	.42*
Entrepreneur	Maintaining supervision over planned changes to improve your unit.	4.3	2.0	2.4	1.5	.53*
Disturbance Handler	Dealing with previously ignored problems (ones which people have known to exist but avoided) which have come to a head.	4.3	2.0	2.4	1.7	.52*
Negotiator	Negotiating with groups outside your command for necessary materials, support, commitments, etc.	4.3	2.1	2.5	1.7	.56*
Figurehead	Answering letters or signing documents as an official representative of your unit.	4.2	1.9	3.1	1.7	.53*
Liaison	Attending outside conferences or meetings.	4.2	1.8	3.2	1.4	.49*
Leader	Participating in EEO activities and responsibilities.	4.1	2.4	1.7	1.1	.33*
Figurehead	Making yourself available to "outsiders" (such as consumers, sponsors, the public) who want to go to "the person in charge."	4.0	2.0	2.6	1.7	.39*
Spokesman	Keeping professional colleagues informed about your unit.	3.9	1.8	2.2	1.6	.59*
Liaison/Monitor	Touring your own command's staff and activities, including field activities	3.6	2.0	1.9	1.5	.58*
Monitor	Gathering information from or about sponsors and consumers.	3.5	2.1	2.2	1.6	.70*
Entrepreneur	Evaluating the outcomes of internal improvement projects.	3.5	2.0	1.7	1.1	.56*
Negotiator	Working with people to see that necessary contracts get negotiated.	3.3	2.4	1.5	1.4	.59*
Figurehead	Escorting and briefing official visitors.	3.2	2.0	1.6	1.2	.33*
Monitor	Monitoring output of formal management information systems, including productivity measures and cost accounting records.	3.2	2.3	1.5	1.3	.63*
Technical Expert	Directing a technical project or subproject.	3.1	2.6	1.8	2.0	.72*
Negotiator	Participating alone or on a team in atypical negotiations with outsiders.	3.0	2.4	1.5	1.2	.70*
Figurehead	Attending business meetings or social gatherings as an official representative of your unit or command.	3.0	2.2	1.5	1.3	.49*
Liaison	Developing new contacts by answering requests for information.	2.8	2.0	1.6	1.5	.61*
Technical Expert	Identifying and solving complex engineering or scientific problems yourself.	2.8	2.4	1.6	1.8	.72*
Negotiator	Handling formal grievances.	2.7	2.7	0.7	0.1	.31*
Liaison	Joining boards, organizations, clubs, or doing public service work which might provide useful work-related contacts	2.3	2.0	1.2	1.1	.68*
Spokesman	Keeping the general public informed about your unit's activities, plans, or capabilities.	1.8	2.0	0.9	1.3	.67*
Negotiator	Negotiating labor-management agreements.	0.6	1.5	0.2	0.5	.51*

[a] Based on an 8-point scale, where 0 = None and 7 = Great Deal.

[b] One of five highest ranked items as to time spent.

*p < .05.

Personal Characteristics Required of Effective Executives

In the civilian questionnaire, executives were presented with a list of 30 characteristics and asked to rate their importance to job effectiveness. These characteristics were identified as important during the interviews. The most important ones concern oral and written communication, listening to others, technical ability, managerial ability, critical thinking, and persuasiveness. The least important concern social relationships with work associates, building a power base, and survival skills.

The 30 items were factor analyzed using varimax rotation. Six major factors emerged, which accounted for 56% of the variance:

1. Interpersonal skills--involves the ability to communicate verbally and in writing, listening skills, flexibility, and persuasiveness.
2. Administrative ability--involves the ability to plan, to process paperwork and other organizational demands, and to manage both time and externally imposed crises.
3. Risk taking ability--includes willingness to take risks, to question directives, and to be achievement-oriented.
4. Awareness of power--refers to survival skills and building a power base.
5. Technical skills--includes technical ability and keeping up-to-date in one's technical specialty.
6. Managerial ability--includes the ability to create an effective work environment for subordinates and to plan and direct the work of an organizational unit.

Training and Development

In conventional job analysis, task and/or ability information is used to infer training needs. In our study, however, we also collected direct opinions from civilian executives and their military superiors about civilian executive training needs. The first important point to emerge was that most respondents felt experiential or on-the-job development was preferable to formal, classroom training. If used at all, it was felt the latter should be employed as a supplement to the former.

Since most executives have relevant academic backgrounds and have worked in a technical area throughout their careers, technical training is not considered especially critical. Refresher training and technical updates, however, especially in areas of rapid technological change, are considered important for keeping up with the state-of-the-art. Instead, the majority of executive training needs were found to be in the areas of general management skills and how the Navy/DoD/Civil Service system functions. In the civilian questionnaire, executives were asked to rate 14 general topics and 30 specific topics as to their importance for prospective Navy civilian executives in terms of training and development needs. Communication and interpersonal-type skills are seen as the most important general subjects. RDT&E management, project management, and the role functions of DoD are seen as the most important specific topics.

Beyond information pertaining to training content and method, an equally important third aspect of training emerged--namely the factors affecting participation in training. Here our findings produced a picture of a system in which there are more factors discouraging than encouraging civilian executives to participate in training. The major deterrent is the pressure to produce; that is, the inability to take time away from the regular job. Other factors include (1) the perception that participation is not related to or is negatively related to promotions or high performance ratings, (2) a fear of being displaced while absent if training requires leaving one's job temporarily, and (3) unwillingness to move if training requires geographical mobility. In addition, there are factors that discourage commands from allowing their personnel to participate. Again, the most important is pressure to produce and the inability to spare employees under conditions of personnel shortages.

Findings Regarding Jobs of Military Executives

In the Introduction it was mentioned that in the process of doing our study, we did indirectly obtain information about the jobs of senior Navy officers in the shore establishment. In general, there was found tremendous similarity between their jobs and the jobs of civilian executives in terms of job content, characteristics, environment, and skills needed. The major difference is that military executives are not as delimited in terms of working within a narrow technical or functional specialty. Consequently, the range of demands is broader, and there is a greater need for them to acquire and rely on specialized staff expertise.

DISCUSSION

An important feature of this particular study was that it constituted a large scale job analysis of a managerial job using a multimethod approach. A number of things were learned from this experience which are applicable to future managerial job analyses. One was that three of the four individual data collection methods proved very useful. The one which did not prove useful was the work diary, simply because executives are too busy to fill it out in sufficient detail. The remaining three methods--interviews, observations, and questionnaires--were not only useful individually but even more so when the data were combined. It did become apparent that information from any single method by itself would give an incomplete picture of the job. For example, the nature and profound impact of the job environment came through most clearly in the interviews. On the other hand, the hectic, fragmented nature of the job came through most clearly in the observations. The survey, in turn, provided quantitative information about the various subtasks of the job. However, since most military job analysis takes a survey approach and breaks the job down into micro task units, we wish to add a note of caution about applying this approach to managerial jobs. Based on our interviews and especially on our observations of executives, we think there are certain aspects of the managerial job which can not be picked up by the conventional task analysis approach. These aspects have to do with the process of management, such as how managers juggle their time, balance competing objectives, cope with a highly fluid environment, etc. The structured survey is better suited to picking up static, job content but misses out on the dynamic aspects of the job. The latter, we are convinced, are just as important as the former.

REFERENCES

Broedling, L. A., & Lau, A. W. Executive Summary: Navy civilian executive study (NPRDC SR 79-10). San Diego: Navy Personnel Research and Development Center, January 1979.

Lau, A. W., Broedling, L. A., Walters, S. K., Newman, A., & Harvey, P. M. The nature of the Navy civilian executive job: Behavior and development (NPRDC TR 79-27). San Diego: Navy Personnel Research and Development Center, July 1979.

McCall, M. W., & Lombardo, M. M. Leadership: Where else can we go? Durham, NC: Duke University Press, 1978.

McCormick, E. J. Job and task analysis. In M. D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Mintzberg, H. The nature of managerial work. New York: Harper and Row, 1973.

HOW THIN DO WE SLICE THE OFFICER JOB TASK BALONEY: A COAST GUARD VIEW

Richard S. Lanterman
United States Coast Guard Headquarters
Washington, D.C.

My purpose in this presentation is to discuss four approaches to Officer Job Task Analysis that the Coast Guard has used and to open for discussion what I see as mistakes and successes in each of them.

Let me walk quickly through our brief experience with Job Task Analysis. Our first excursion was a contract study of the job of the Junior Officer. The aim was to validate the curriculum at the Coast Guard Academy. The survey and subsequent analysis showed the Academy curriculum was reasonably appropriate to the first five years of a Coast Guard officer's career, but the survey could have been better. Some of the mistakes we made were:

1. The survey was too long.

2. Many of the tasks were vaguely phased.

3. Many of the tasks were not time rateable.

As an example of that last point, one of the tasks asked was: "Relative to all other tasks you do, how much time do you spend assessing your ability to learn the language of the 'old Salt'?"

We then stumbled into the study of an officer specialty, in spite of the fact that: 1) The CG concept is that all officers are generalists, and 2) there are only about 5,000 officers in the entire force. There is obviously some danger in trying to slice the baloney too thin.

Next, we had a good idea. (I've kept using the editorial "we" here hoping that you will assume that someone else shares the blame for my mistakes. No one does.) This study was reported at the last MTA. We asked all officers and warrant officers in the Coast Guard to describe their jobs by indicating the amount of training they felt a new incumbent would need to do the job well. The "tasks" or training elements available to describe the billet (job) came from a catalog of courses and were those that officers might be exposed to in some kind of postgraduate or advanced technical training. The process was easy for respondents to handle although we made two big mistakes. One was to ignore, through oversight, Aeronautical Engineering as a subject area. The other was to ask the officers to describe their own skill level in each of the 600 plus courses, after they had described their billet training needs. This was useful in giving us a picture of the discrepancy between resources and requirements, but it immensely increased the time of administration. All officers apparently think they know at least something about everything. But, since people love to tell you about themselves, most persevered . At least one Flag officer reported taking seven hours to complete the survey.

The major findings of the survey were: 1) with a few exceptions, there is more than enough talent available to satisfy the billet training requirements, and 2) an overwhelming part of all officers' jobs involve communication skills and, to a slightly lesser extent, management skills. I recommend this approach of describing jobs in terms of curricular (training) requirements. It is easy for incumbents and may save some of the heartburn involved in mapping the usual job tasks back through skills and knowledges to training objectives.

Finally, our current effort in officer job task analysis takes off from the finding of the preceeding study that, at least for Coast Guard officers, the technical skill requirements of the job are very much subordinate to the communication and management skills. The main place we teach these skills after commissioning is in our Leadership and Management schools in Yorktown, VA and Petaluma, CA. The staff of the Yorktown school approached us with the idea of doing a training task analysis of leadership, management, and administrative requirements of supervisory jobs, not just at the officer level, but for enlisted supervisors as well. This with the heretical assumption that officers are not the exclusive owners of "Leadership." The inventory was developed primarily by Chief Petty Officer Steven Wehrenberg of the Yorktown school staff after many field interviews with supervisors at all levels, and, of course, was based on his perception as an instructor as to what should be taught.

There are only 284 tasks in the inventory, divided into eight duty categories. The items seem to be well conceived. They are similar to those in an Air Force OMC survey which Jerry Baruckey was going to report on yesterday and many of the tasks look like those in the Supervision and Management section of the U. S. Army MILPERCEN's officer occupational survey pilot project. How could they not? The difference is that we stop there, on the conviction that supervision and management are the job of the supervisor and that technical subject matter skills (e.g., navigation, oil spill analysis, etc.) while necessary, can usually be assumed to be present. In any event, these are subservient to the management aspects of the jobs.

To Summarize:

1) At least for Coast Guard officers, there is a need for a large amount communication skills and management skills.

2) Non-commissioned officers are not different in kind from other officers.

3) There are useful ways of describing officers and other "non-performance" jobs.

ANALYSIS FOR THE AUSTRALIAN ARMY OFFICER TRAINING AND EDUCATION CURRICULUM

Major Paul F. Routh

Headquarters Training Command PO Box 39
Darlinghurst Sydney 2010 Australia

The views expressed in this paper are the author's own and should not be constructed as Australian Department of Defence (Army Office) opinion or policy

ANALYSIS FOR THE AUSTRALIAN ARMY OFFICER

TRAINING AND EDUCATION CURRICULUM

Major Paul F. Routh

Headquarters Training Command PO Box 39

Darlinghurst SYDNEY 2010 AUSTRALIA

INTRODUCTION

The Australian Army has had considerable experience in Occupational Analysis for soldiers and Non Commissioned Officers particularly in the area of skill training. A marked improvement in the efficiency and effectiveness of individual training has been achieved with the implementation of the Army Training System, a system similar to the US Interservice Procedures for Instructional Systems Development (Branson et al 1975). Comprehensive Occupational Data Analysis Programmes (CODAP) are used in the analysis phase of the System which ensures that the analysis of training needs is based directly on the tasks performed rather than on opinions of what should be taught.

However problems arose in the application of similar procedures to the determination of officer education and training requirements. One problem was the need to identify the general competencies required to perform a range of tasks now and in the future. Another, was that current curriculum design for general education tended to concentrate on the identification of the content of various subject or study areas while traditional job analysis for training focused upon duties and tasks of specific employments. Studies carried out in civilian organizations have indicated that similar problems have been experienced in the design of management training and development.

Due to a period of general stability during the last decade the Army has been able to reorient its policies and priorities. As a result, various studies have been conducted which have investigated aspects of the Army's organization, administration and use of resources. One of these studies was the Regular Officer Development Committee (RODC) project. This project investigated the requirements for officer development in the 1980s and beyond. The Project Report (1978) contained a number of recommendations for improving officer development and employment in the Army and emphasised the following:

a. There was a strong requirement to improve the existing officer education and training to overcome a number of deficiencies evident with officer development, employment and performance.

b. Officer education and training should be such that the various planned learning experiences should be appropriate to rank and career progression.

c. The recognition that an officer career span of 20 years or more justifies investment of resources into officer occupational education as well as vocational training.

Occupational Education may be described as that which focuses on competencies that go beyond immediate task mastery. Such general competencies support the performance of a range of tasks now and in the future where it is likely that a proportion of tasks will be difficult to predict.

As a result of these findings the project which is the subject of this paper was then initiated. The requirement was to prepare a curriculum plan for the command and staff training of lieutenant colonels and the staff and operations training for junior officers that was not special to a particular Corps or Branch of the Service such as Infantry or Signals. However it was realized that such training could not be divorced from other aspects of officer development and that the analysis would have to cover all aspects of military education and training from pre-commissioning to lieutenant colonel.

This paper describes the way the initial phase of the study was conducted, and describes the results that have been obtained to date.

THE APPROACH

It was resolved that the steps of the Army Training System for Analysis and Design would form the basis of the methodology. However variations for officer training and education would be required.

An earlier study of the Australian Army logistic training requirement had proved the value of functional analysis. Using a pattern similar to that employed by the US Army in its Logistic Basic Functional Structure (1975) a vast array of tasks were organized into a reasonably simple structure. Job analysis then concentrated on the involvement of the various ranks and employment groups.

This same sequence was then applied to the whole range of officer employment. Job analysis data on several officer positions were available. However inventories were not common, and sub-populations in the samples were small and 'time spent' was the only parameter investigated.

Occupational Education was seen as a legitimate goal for the study. It was recognized that not all tasks of the future could be predicted and there was a need to identify the general competencies that would assist in this regard. Accordingly the following curriculum items were assembled in addition to the tasks that had resulted from the functional and job analyses:

a. generalized skills, for example, problem solving skills or communication skills;

b. general concepts which would have wide application, for example the concept of a command and control system; and

c. bodies of knowledge which could be conveniently grouped such as the roles, organization and tasks of an infantry battalion.

Panels of experienced officers were used extensively for the collection and organization of the items. The panels were briefed on an approximate curriculum structure. This structure allowed for the organization of the total project into manageable sub projects as well as providing a means of collating the curriculum items that had been identified. The curriculum structure is represented diagrammatically in Figure 1.

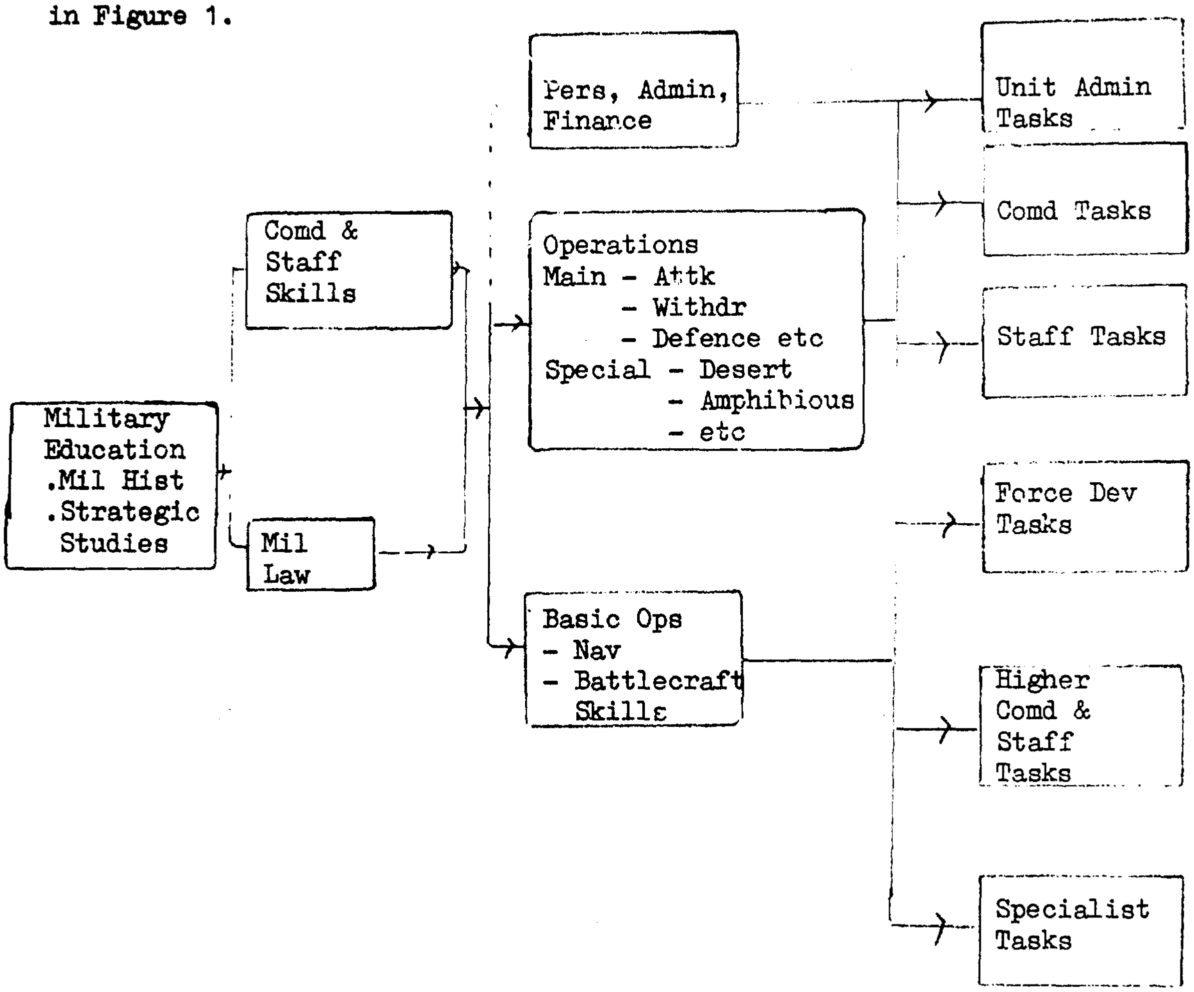


Figure 1: Curriculum Structure for Officer Training and Military Education

The structure shown had the effect of limiting the scope of the lists prepared in each of the general sections. Items could only be included that would support a range of tasks that had already been identified.

Estimates of priority were made for each curriculum item to allow for contraints in training time and resources to be taken into account at a later stage. These estimates took the form of an importance rating and a rating of proficiency required. Approximate time to learn was also estimated based on experience from the teaching of similar material.

To assist panel members in the estimate of importance of the various curriculum items, reference was made to a Military Education Topics Study (1977). This study involved a wide survey to determine the opinion of various officer ranks in regard to the priority of traditional military education subject topics. Respondents rated each topic in order of priority of 'needed on job' and 'need to know'. A sample page showing results of this study is at Annex A.

RESULTS

As a result of its analysis, the study team produced a comprehensive curriculum list for all-corps officer training ranging from pre-commissioning to command and staff training at lieutenant colonel level. Sample pages of this key document are included as Annex B. Significant characteristics of the list are as follows:

a. The list shows what to teach and when to teach it and is therefore the basis for the design and development of instruction. Problems occur in the management of training when each training institution responsible for a stage in the progression tends to teach the same material or use the same exercises or tests. The list should show training progression. However in some cases this progression is not clear and training designers must then be alerted to clarify the boundary between what is planned for prior and subsequent training.

b. The list allows for the application of constraints. Cost-benefit estimates on each item can be taken into account when deciding what to teach or delete or to teach at a lower standard of proficiency.

c. The list provides information that assists in the grouping of the various learning activities required. Courses can then be given appropriate titles and general objectives.

d. The list allows for quick changes in emphasis. For example, if national defence emphasis changes from counter insurgency to amphibious operations, the curriculum items that have not been taught in detail can be given greater emphasis.

e. Training for special groups of officers can be determined based on their planned employment in the generalist area. This is particularly useful for specialist officers and the Army Reserve.

f. Special to Corps or Branch of the Service training need not duplicate what has already been mastered in the all-corps training list.

It is probable that insufficient information will sometimes exist for the preparation of a learning objective, exercise or test. There is also the possibility that some general competencies or tasks may have been overlooked. However, inventories of such competencies are reasonably valid based as they are on recent combat experience and annual training exercises. It is intended that such limitations should require training personnel to conduct further analysis of the inadequate set of curriculum items. Such an investigation would usually be delegated to a number of training institutions before being reviewed and implemented by a central agency.

The list was been submitted for approval as the basic document to be used for the design and management of officer training in the Australian Army.

CONCLUSION

Special features of officer training and education have justified the inclusion of curriculum items of a more general nature that may require application in the uncertain environment of the future. The management of such learning has required the preparation of a comprehensive curriculum list which should form the basis for control of the progression of training and education of officers.

The philosophy of the approach taken for the training and education of Australian Army Officers does not differ in principle from that of the Army Training System as applied to the training of normal military skills. Effective and efficient training is still the primary goal of the procedures that have been developed.

REFERENCES

Auctralian Army ... Report of the Regular Officer Development Committee, Defence Printing Establishment 1978.

... A Survey of Educational Requirements as Seen by Australian Army Officers, report by Military Employments Research and Information Team, 1977.

Branson, R. K., Raynor, G.T., Cox, J.L., Furman, J.P., King, F.T. and Hannum, W.H. Interservice Procedures for Instructional Systems Development, (5 Vols.)
(TRADOC Pam 350-30 and NAVEDTRA 106A)
US Army Training and Doctrine Command, Ft Monroe, Va: August 1975.

Morsh, J.E. Survey of Air Force Officer Management Activities and Evaluation of Professional Military Education Requirements, USAF Human Resources Laboratory, Personnel Research Division, 1969.

US Army - Logistics Basic Functional Structure,
(Army Regulation No 700-126). Headquarters Department of the Army Washington DC, 3 March 1975.

Annexes: A. Military Education Topics Survey 1977 - Sample of Results

B. Sample Pages From All Corps Officer Training List

MILITARY EDUCATION TOPICS SURVEY 1977 - SAMPLE OF RESULTS

SUMMARY

The opinion of 473 officers was sought on 122 Military Education Topics. Respondents rated each topic in. regard to 'needed on job' (NOJ) and 'need to know' (NTK). The 122 topics were condensed into 14 significant study areas. The survey was based on a similar study, conducted for the USAF (Morsh 69).

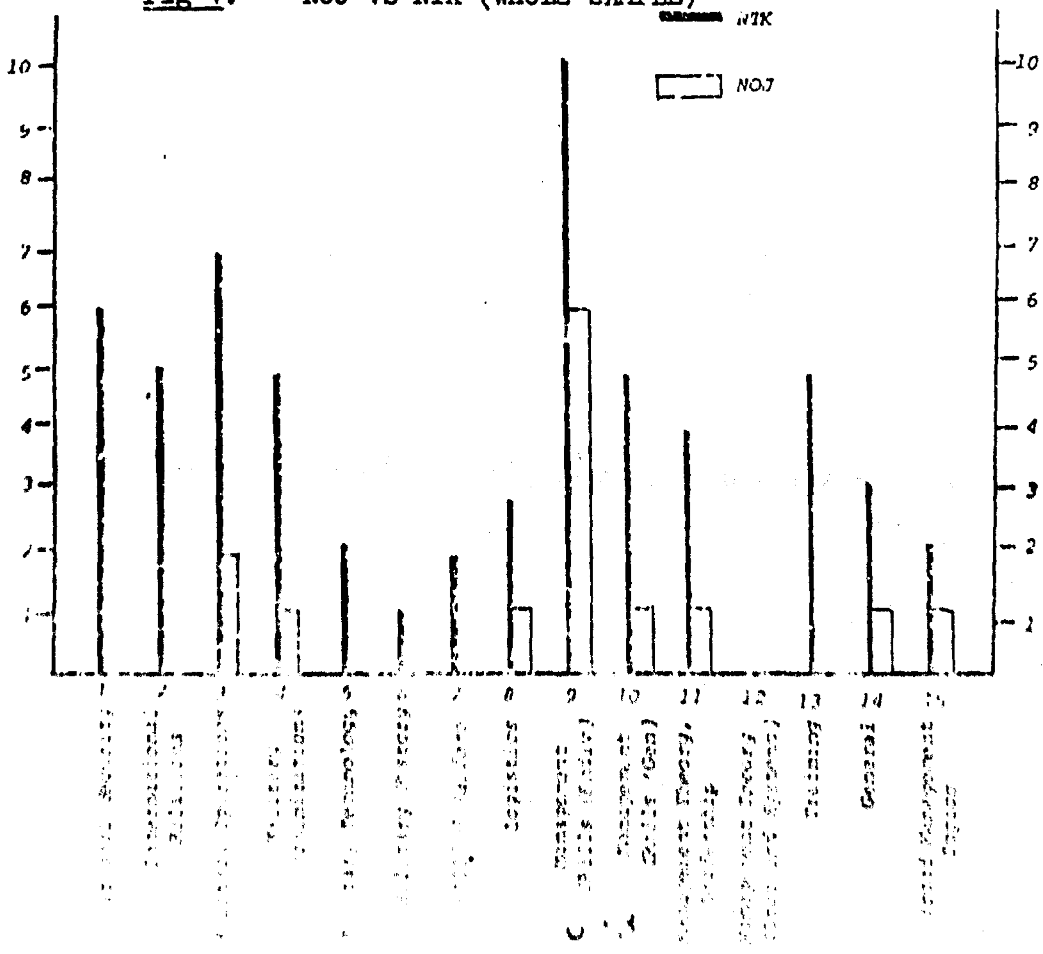
RESULTS

When mean ratings are related to the rating scale, a mean of 4.0 or higher indicates a 'substantial need', and topics with mean ratings in this category were considered as a basis for compiling the following results.

NTK vs NOJ - Whole Sample

A summary of the number of significant mean Need to Know responses (i.e. mean 4.0) compared to the number of significant Need on Job responses, for each Topic Heading, appears in Figure 1 below.

Fig 1. NOJ VS NTK (WHOLE SAMPLE)



SAMPLE PAGES FROM ALL-CORPS OFFICER TRAINING LIST

General Military Education

GME 1	Military History
GME 2	Strategic Studies

Command and Staff Skills

CS1	Command and Control
CS2	Command and Control Systems
CS3	Defence Force and Civil Organizations
CS4	Training
CS5	Leadership
CS6	Communication Skills
CS7	Military Law

Operations

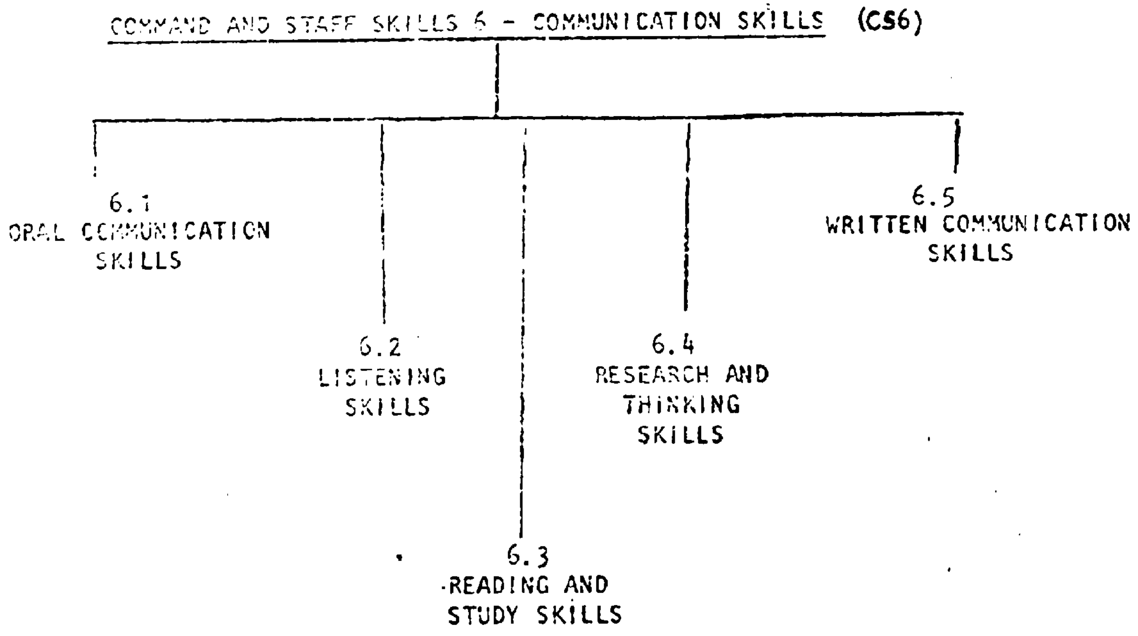
OPS1	Basic Operations
OPS2	Main Types of Operations
OPS3	Operations Under Special Conditions

Administration

UADM	Unit Administration
AADM	Defence Administration
OADM	Operations Administration

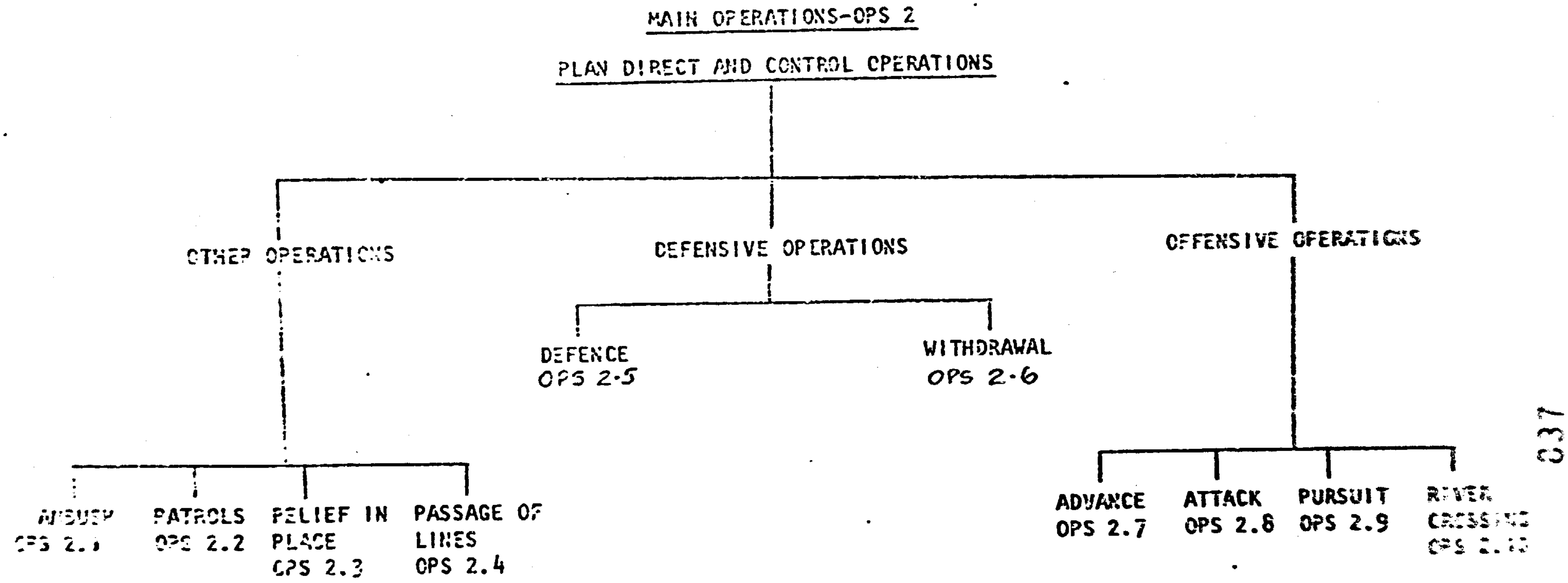
Miscellaneous

DR	Drill
PT	Physical Training


COMMAND AND STAFF SKILLS 6 - COMMUNICATION SKILLS (CS6)
6.1 ORAL COMMUNICATION SKILLS
6.2 LISTENING SKILLS
6.3 READING AND STUDY SKILLS
6.4 RESEARCH AND THINKING SKILLS
6.5 WRITTEN COMMUNICATION SKILLS

Serial	General Statement of Tasks	RMC of A			SOC				IOC				C & SC			
		Level	Imp	Time Est	Level	Imp	Time Est (Formal)	Time Est (Ext)	Level	Imp	Time Est (Formal)	Time Est (Ext)	Level	Imp	Time Est (Formal)	Time Est (Ext)
	RESEARCH AND THINKING SKILLS CS6.4															
	General Statement: In practical oral and written tasks use a methodical, thorough research process to gather, evaluate, interpret and use data (fact, opinion, argument criticism).															
6.4.1	Use a methodical, thorough research process to gather, evaluate, interpret and use data. a. identify a problem from data. b. identify central issues and underlying assumptions c. Formulate and state hypothesis d. recognize logical implications of hypothesis			e. decide additional data needed f. gather and record data												

Note: 15. Instruction. To be practised in out of formal training time and other communication skills practical exercises.


MAIN OPERATIONS-OPS 2
PLAN DIRECT AND CONTROL OPERATIONS
OTHER OPERATIONS
DEFENSIVE OPERATIONS
OFFENSIVE OPERATIONS
AMBUSH OPS 2.1
PATROLS OPS 2.2
RELIEF IN PLACE OPS 2.3
PASSAGE OF LINES OPS 2.4
DEFENCE OPS 2·5
WITHDRAWAL OPS 2·6
ADVANCE OPS 2.7
ATTACK OPS 2.8
PURSUIT OPS 2.9
RIVER CROSSING OPS 2.10

		RMC of A			SOC				IOC				C & SC			
Serial	General Statement of Tasks	Level	Imp	Time Est	Level	Imp	Time Est (Formal)	Time Est (Ext)	Level	Imp	Time Est (Formal)	Time Est (Ext)	Level	Imp	Time Est (Formal)	Time Est (Ext)
.6	Arrange services to maintain the rate of advance.															
.7	Determine and implement bypassing policy.															
.8	Plan and coordinate fire support for the advance.															
.9	Plan and coordinate air support to cover the advance.															
.10	Produce an obstacle clearance plan.															
.11	Maintain liaison with flanking units.															
.12	Conduct an advance.															
	OPS2.8															
2.8	Plan, Direct and Control an Attack															
2.8.1	Command an attack.															
.2	Select:															
	.1 concentration areas;															
	.2 assembly areas;															
	.3 approaches;															
	.4 FUP;															

OFFICER JOB ANALYSIS: IS THE ENLISTED MODEL APPROPRIATE?

Hendrick W. Ruck

Air Force Human Resources Laboratory
Brooks AFB, Texas 78235

INTRODUCTION

Let me begin by saying that I am honored to have been asked to speak at this symposium. I must admit that I have not performed research on officer job analysis while at AFHRL. In fact, research on officer job analysis has been rather dormant for the past five years (Mayo, Nance, & Shigekawa, 1975). What, then qualifies me as an "expert?" Two words--training and operations. I have been involved in, together with several other psychologists, research on the development of training requirements using occupational survey data (Christal, 1970; Mead, 1975; Mial & Christal, 1974; Stacy, Thompson, & Thomson, 1977; Ruck & Birdlebough, 1977; Ruck, Dineen, & Cunningham, 1977; Ruck, Thompson, Thomson, 1978). Although the use of occupational survey data in determining training requirements has long been of interest in the Air Force, only recently has it become important in the other services. Now for operations. Several years ago, when working at the USAF Occupational Measurement Center, I was responsible for developing the second operational officer job inventory, and editing the survey report (Tauscher, Ruck, DiTullio, & Stephenson, 1977). I have advised on most of the subsequent officer occupational surveys. So much for me. Now for the symposium: How thin can officer job analysis baloney be sliced? That is a very good question.

The military services have been performing operational occupational analyses of enlisted specialties for about a decade. Most of the enlisted Air Force Specialties (AFS), Military Occupational Specialties (MOS), and Navy Ratings have been studied and results provided to personnel, training and manpower managers. The success of the enlisted programs has resulted in considerable interest within the using communities in performing occupational analyses for officer specialties. This interest is underscored by the establishment of a section to perform officer surveys in the USAF Occupational Measurement Center, and, more recently, the initiation of the Army's program to measure officer jobs to establish training requirements. The purpose of this paper is to explore some of the similarities and differences that may be found in analyzing officer jobs compared with analyzing enlisted jobs.

ASSUMPTIONS

Prior to delineating some of the assumptions made in performing enlisted studies, it is appropriate to examine the term occupational analysis. What is it? Shartle (1959) has defined an occupation as a group of jobs which are similar and are usually classified together in different agencies. He further defined a job as a group of similar positions, and a position as a group of tasks. Occupational survey data are used for analyzing at both the occupation and job levels for enlisted personnel.

Let us take a look at some of the key assumptions made when using occupational survey methodology to measure jobs and occupations. Although these assumptions are mine, I am indebted to some of the early researchers (Archer & Fruchter, 1963; Morsh & Archer, 1967) for my inspiration. Table 1 lists several of the major assumptions made in applying occupational survey methodology.

Time and space limitations do not allow a full discussion of each of the assumptions, so only a brief comment will be made for each. The first two assumptions, there are common skills, knowledge, tasks, and activities, and the fourth assumption, jobs exist, are required to allow job inventories or task lists (and associated background questions) to be of reasonable length. A reasonable number of tasks statements that are meaningful, clear, and concise should be derivable for a specialty. If there is little commonality within a specialty, useful task lists would be exceedingly difficult, if not impossible, to construct. The third assumption, there is a common language or terminclogy, is required to apply the occupational survey methodology successfully, since job incumbents provide responses to the surveys. If terminology were ambiguous or conflicting, self-report responses may be quite difficult to interpret. The fifth assumption, careers exist within specialties, is, perhaps least important. It allows inventory developers tc work with relatively few "experts" in developing the job inventory and provides the possibility of performing career path analyses.

TABLE 1

Assumptions Made When Using
Occupational Survey/Analysis Methodology

<u>Occupation</u>

Common underlying skills/knowledge exist
Common tasks/activities exist
Common language/terminology exists
Jobs exist, not just positions
Incumbents rotato among jobs during career

<u>User</u>

Occupational structure exists
Trainers need task information

The assumption that an occupational structure exists is critical for two reasons. Without such structure, the survey instrument would be written in such broad terms as to be, perhaps, useless. Also, without such structure, or, at least, the intent of such structure, personnel and manpower managers would be lost as valuable customers of occupational survey data.

The final assumption, trainers need task information, is critical. In highly proceduralized specialties, such as pilot or missile launch officer, so much of the critical portion of the job has been analyzed and documented that occupational survey data would add very little to the trainer's information base. However, in less well documented specialties, there may be a need for survey data.

It is important to note that technologies other than occupational analysis have been developed at AFHRL to evaluate officer grade requirements (Stacy, Matthews, & Hazel, 1975) and to establish officer education requirements (Goody, 1977) primarily because survey data would not provide appropriate information. Before examining the assumptions as they relate to officer specialties, let us take a look at the officer personnel classification structure as it exists in the Air Force.

OFFICER CLASSIFICATION AND ANALYSIS

There are 95,200 officers on active duty in the Air Force at present. The officers serve in 227 different specialties which combine to make 53 utilization fields. These utilization fields are combined into 22 career areas. Table 2 presents the approximate active officer strength in a number of key utilization fields. Note that the terminology used in defining officer specialties, utilization fields, and career areas, suggests a certain impermanence of association between the job incumbent and his or her specialty.

TABLE 2

Approximate Officer Manning in
Selected Utilization Fields

	N	% Total
Pilot/Navigator	30,000	29.7
Missile Operations	3,400	3.3
Special Duty/Commander/Director	10,200	9.9
Medical/Dental Veterinarian	8,300	8.0
Nurse	4,100	4.0
Judge Advocate	1,200	1.1
Chaplain	900	.8
Unclassified	1,500	1.5
	59,600	58.3

Now, let us review the selected utilization fields (making up almost 60% of the officer force) to determine how they measure upon the key assumptions concerning application of the occupational survey approach (Table 3).

Note that only one of the utilization fields satisfies all seven assumptions, and five of the fields satisfy four or fewer of the assumptions. The implication is that occupational survey methodology is not directly appropriate in utilization fields containing more than half of the active duty officers.

At this point, one may argue that I have set up an unrealistic scenario; that no one would seriously consider using the occupational survey approach for all officer fields, and that the fields I have chosen are too broad. The first argument has been addressed--there is serious interest in performing large-scale wide-range analyses of most officer jobs in at least two of the services. Let me respond to the second argument first by reviewing studies that have been performed in the Air Force, and then by comparing the analysis of specialties, rather than utilization fields, with the assumptions delineated earlier.

TABLE 3

Assumptions Versus Utilization Fields

	Pilot/Navigator	Missile Operation	Special Duty	Medical/Dental/Veterinarian	Nurse	Judge Advocate	Chaplain	Unclassified
Common Skills/Knowledge	X	X		X	X	X	X	
Common Tasks/Activities	X	X		X	X	X	X	
Common Language	X	X		X	X	X	X	
Jobs Exist	X	X		X	X	X	X	
Job Rotation (Career)				X	X	X		
Occupation Structure				X	X	X		
Trainers' Needs					X			

Since 1963, 37 different officer occupational surveys have been conducted in the Air Force. Of the 37 studies, 31 (84%) were conducted on utilization fields. Reviewing these 31 studies, the prime purpose of each was to examine occupational structure. However, those studies done before 1972 (N=19) resulted largely in the finding that officers perform similar tasks within the same utilization field, regardless of rank. Mayo et al (1975) suggest that these findings are a result of task statements being managerially oriented and broadly stated, similar to Hemphill's (1960) instruments. Inventories subsequent to 1972 (N=13) have been more task oriented and have produced more specific occupational structures. Virtually all of the officer occupational analysis since 1975 has been performed in the operational environment. Of the seven occupational surveys conducted since 1975, four have been performed at the specialty level and have been developed primarily for training requirements development. It is important to note that these four specialties, signals intelligence, social actions, space systems, and weapons control, are each highly technical specialties, requiring extensive detailed job knowledge and having no enlisted counterpart. Thus, they are more like enlisted specialties than traditional officer specialties.

Let us compare officer specialties with the assumptions listed in Table 1. Table 4 lists four specialties which have been studied and a selected sample of other specialties from the spectrum of officer specialties. The first four specialties listed are currently being or have recently been studied. For three of those specialties, the occupational survey approach may have been only marginally cost effective due to the small size of the specialty populations. Also note that only one of the specialties met all my assumptions in applying occupational survey methodology.

TABLE 4

Assumptions Versus Specialties

	Signals Intelligence	Social Actions	Space Systems	Weapons Control	Transport Pilot	Weather Officer	Personnel Officer
Common Skill/Knowledge	X	X	X	X	X	X	X
Common Tasks/Activities	X	X	X	X	X	X	X
Common Language	X	X	X	X	X	X	X
Jobs Exist	X	X	X	X	X	X	X
Job Rotation (Career)						X	X
Occupation Structure						X	X
Trainers' Needs	X	X	X	X		X	
Population N	304	53	350	1298	2241	1028	1525

Tables 3 and 4 depict a gloomy picture because so few assumptions are actually met. Should we abandon the occupational survey method in dealing with officers? No, the occupational survey approach may be useful and effective in some of utilization fields or specialties. For example, it could be used to measure jobs in the transportation, personnel, administration, supply, aircraft maintenance, communications maintenance, or security police utilization fields (to cite a few); but, to what end? Is occupational structure important? I do not know.

An important study might be conducted on the flow of officers to and from those fields. There may be more inter-field career movement than intra-field. If so, occupational analysis would not be nearly as useful as some other measure such as "utilization flow analysis." Suppose officers were found to progress in a career within a utilization field. In that case, training managers may want occupational survey data. The question is, why? Traditionally, officer technical training in the Air Force covers the enlisted job in detail and adds some management type training. Survey data gathered from officers would probably not substantiate such a training approach because the officer performs management tasks that require understanding, but not performance, of the enlisted job. Officer specialties that are highly technical, and require job training, and, hence, training requirements analysis, do exist. Our experience in developing training requirements for enlisted technical analyses shows that curriculum development personnel need more information than percent performing and a relative time spent index for tasks. Scales of relative learning difficulty and recommended entry-level training emphasis have been developed for use in enlisted studies. Such scales have not been adequately researched for officer job analysis. There is a clear need to develop scales that are appropriate for officer tasks if we are going to measure officer jobs for training requirements.

You may ask, then, what about management-type training? Occupational surveys focused upon specific tasks do not usually cover management skill and knowledge requirements in such a way as to help trainers or educators very much. Other, broad-brush approaches (Morsh, 1969) may be more suitable for establishing management training requirements. Level of responsibility and decision complexity, for example, are broader characteristics worthy of attention when officer jobs are described.

CONCLUSIONS

After painting such a gloomy picture about the appropriateness of the enlisted model, one might expect a recommendation to forsake occupational survey methodology in the analysis of officer jobs. I do not make such a recommendation. What I do recommend is caution and prudence. The blind application of a methodology to officer specialties because it worked very well in enlisted specialties would result in a very large expenditure of resources for, perhaps, very little return. Occupational surveys can be quite effective in producing meaningful data for making management decisions about officer specialties. However, a user of the data with clear-cut requirements should be involved from the start, and most of the assumptions listed in Table 1 should be satisfied. Otherwise, alternative analytic methods, such as free format reporting (Fruchter, Morin & Archer, 1963), policy capturing on education/experience profiles (Goody, 1977), course content surveys (Morsh, 1969), or standardized measures (Mitchell & McCormick, 1976) may prove considerably more productive.

In studying officer specialties, the analyst must "work harder (Mayo, et al, p. 76, 1975)" because the work itself is considerably more complex.

REFERENCES

Archer, W.B., & Fruchter, D.A. The construction, review, and administration of Air Force job inventories . PRL-TDR-63-21, AD-426-755. Lackland AFB, TX: August 1963.

Christal, R.E. Implications of Air Force occupational research for curriculum design. In B.B. Smith & J. Moss, Jr. (Eds.), Report of a seminar: Process and techniques of vocational curriculum development. Minnesota Research Coordination Unit for Vocational Education, University of Minnesota, Minneapolis, MN, April 1970.

Fruchter, B., Morin, R.E., & Archer, W.B. Efficiency of the open-ended inventory in eliciting task statements. PRL-TDR-63-8, AD-418 980. Lackland AFB, TX: March 1963.

Goody, K. Matching job education requirements of a variety of officer specialties with the educational attainments of potential incumbents . AFHRL-TR-77-44, AD-A050 826. Brooks AFB, TX: Air Force Human Resources Laboratory, Occupation and Manpower Research Division, August 1977.

Hemphill, J.K. Dimensions of executive positions: A study of the basic characteristics of the positions of ninety-three business executives . Columbus, Bureau of Business Research, College of Commerce and Administration, Ohio State University, 1960.

Mayo, C.C., Nance, D.M., & Shigekawa, L. Evaluation of the job inventory approach in analyzing USAF officer utilization fields. AFHRL-TR-75-22, AD-A014 800. Lackland AFB, TX: Air Force Human Resources Laboratory, Occupational and Manpower Research Division, June 1975.

Mead, D.F. Determining training priorities for job tasks. Paper presented at 17th annual conference of the Military Testing Association, Indianapolis, Indiana, September 1975.

Mial, R.P., & Christal, R.E. The determination of training priority for vocational tasks. Proceedings, psychology in the Air Force symposium. USAF Academy, April 1974.

Mitchell, J.L., & McCormick, E.J. Professional and managerial position questionnaire. Department of Psychological Sciences, Purdue University, West Lafayette, Indiana, Purdue Research Foundation, 1976.

Morsh, J.E. Survey of Air Force officer management activities and evaluation of professional military education requirements. AFHRL-TR-69-38, AD-705-574. Lackland AFB, TX: Air Force Human Resources Laboratory, Personnel Research Division, December 1969.

Morsh, J.E., & Archer, W.B. Procedural guide for conducting occupational surveys in the United States Air Force . PRL-TR-67-11, AD-664-036. Lackland AFB, TX: September 1967.

Ruck, H.W., Birdlebough, M.W. An innovation in identifying Air Force qualitative training requirements. Paper presented at the 19th Annual Conference of the Military Testing Association, San Antonio, TX, October 1977.

Ruck, H.W., Dineen, R.T., & Cunningham, C.C. Applying occupational survey data in instructional systems development. Paper presented at the 19th Annual Conference of the Military Testing Association, San Antonio, TX, October 1977.

Ruck, H.W., Thompson, N.A., Thomson, D.C. The collection and prediction of training emphasis ratings for curriculum development. Paper presented at 20th Annual Conference of the Military Testing Association, Oklahoma City, Oklahoma, October 1978.

Shartle, C.L. Occupational information its development and application. Englewood Cliffs, N.J., Prentice-Hall, 1959.

Stacy, W.J., Matthews, G.N., Hazel, J.T. Determination of officer grade requirements by management engineering teams. AFHRL-TR-75-80, AD-A025 309. Lackland AFB, TX: Air Force Human Resources Laboratory, Occupational and Manpower Research Division, December 1975.

Stacy, W.J., Thompson, N.A., & Thomson, D.C. Occupational task factors for instructional systems development. Paper presented at 19th Annual Conference of the Military Testing Association, San Antonio, TX, October 1977.

Tauscher, L.J., Ruck, H.W., DiTullio, P.N., & Stephenson, S.D. Security police officer and security police staff officer utilizations fields. Occupational Survey Report, AFDT 90-81XX-255. Lackland AFB, TX: Occupational Survey Branch, USAF Occupational Measurement Center, November 1977.

OPERATIONAL DATA + RESEARCH = HAMBURGER OR FILET MIGNON?

Sally J. Van Nostrand

US Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, Virginia 22333

INTRODUCTION

The US Army Training and Doctrine Command (TRADOC) requires redesign of all officer training and education to "train for the job." This is difficult as precise job descriptions of officer positions do not exist. In 1978 TRADOC initiated a five year program to adapt and further develop the methodology currently used for analyzing enlisted jobs to generate information about officer positions. The enlisted methodology, now in use in slightly variant forms in all the uniformed services, is based on task inventory ratings by job incumbents and supervisors. The ratings and background data are analyzed through the Comprehensive Occupational Data Analysis Programs (CODAP) computer software system.

As a pilot effort in the five-year program, one warrant officer specialty and four company grade officer specialties are being surveyed independently during calendar year 1979. There are several reasons to expect that the effectiveness of each survey may be different and that none of the pilot surveys by themselves will be completely satisfactory:

a. The proponent schools have gathered the data for writing task (activity[1]) statements in differing ways.

b. The activities are written at many levels of abstraction (specificity).

c. Officer jobs have never before been analyzed using a task inventory rating in an operational setting.

d. Officer jobs tend to be composed extensively of managerial and supervisory duties which are not easily reduced to specific, useful activity statements.

[1] Officer surveys do not use the word "task." It was felt that officers might feel more comfortable rating activities; task seems to imply a procedural duty rather than a cognitive process (often referred to as "soft skills"). One tongue-in-cheek reason, "as the items do not fit the task definitions, they had to be given another name." The activities do not fit the standard task definition criteria. However, TRADOC is presently working on new task definitione and analyzing procedures which will, in the future, allow both procedural tasks and soft skills to be called tasks.

As the Army has never surveyed officers with task inventories and the other services have conducted only special-purpose task list surveys, there is no historical information on the applicability of the method for officer jobs. Neither is there an assessment of the validity of the traditional scales.

It was decided early in the planning stage that data would be gathered for the Relative Time Spent, Task Learning Difficulty, Consequences of Inadequate Performance and Immediacy of Task Performance factors. These four factors are used in the four-factor model for establishing training priority. Previous Army research has found the four-factor model effective for establishing training priorities in enlisted specialties (Gilbert, et. al., 1978; Siebold, 1979). The efficacy of the model for officer jobs has not yet been assessed.

Several writers have suggested that differing types of job or intended uses of data require different scales (Cunningham and Drewes, 1978; Mitchell, 1978a; Siebold and Waldkoetter, 1978; Van Nostrand and Wallis, 1978). Officer jobs are more nearly managerial jobs, especially at the field grade level, than are the majority of enlisted jobs. Relative Time Spent may not be an appropriate measure of managerial tasks. Therefore, it was also decided that additional scales would be tested during the pilot program. An adaptation of the Hemphill scale (Hemphill, 1960), Part of Position, will be administered to some of the incumbents. Supervisors will be given the Training Emphasis scale developed by the US Air Force (Ruck, 1978). All scales are anchored at each point (Mitchell, 1978) and each is a 7-point scale.

The number of officers in each specialty varies widely. The warrant officer specialty has only 60 incumbents, and the duties within the specialty are different enough that some are trained at the Missle and Munition Management School and the others are trained at the Field Artillery School. The four commissioned officer specialties range from less than two hundred incumbents to more than 4,000. Since some of the specialties have few incumbents, all the scales will not be tested in all specialties. A table showing specialties, scales and approximate number of surveys distributed is at Appendix A.

After these preliminary decisions had been made, the Army Research Institute (ARI) was requested to evaluate the pilot program. Devising an evaluation plan for an on-going, operational program proved to be a challenge. There were both pluses and minuses to consider and a number of trade-off decisions to be made. ARI first conceptualized the plan as it would have been if all the preliminary decisions had been ours to make, as well as all the resources - the filet mignon approach. The final plan uses the positive aspects of being part of a highly-visable, well-supported program, and attempts to overcome negative aspects - the hamburger approach.

THE FILET MIGNON APPROACH

In the complete filet mignon approach we compare several data collection methods, of which the task inventory survey is only one. Other than comparison with other methods we would specify the following objectives:

a. Determine an appropriate level of specificity. Some officers would receive surveys with tasks written according to the standard task definition. Other officers occupying the same duty positions would be given lists written at a more general level.

b. Determine whether ordering of the tasks produces bias. Administer the surveys with duty areas and the tasks within duty areas arranged differently. In addition, survey samples of officers who respond to one duty area only.

c. Determine maximum length possible for reliable results (before fatigue or boredom factors interfere).

d. Determine a minimum number of items within each duty area which produces the "same" results.

e. Experiment with many types of scales, including the number of points per scale for each.

f. Determine test/retest reliability.

g. Compare all of the above among combat arms, combat support, and combat service support specialties.

THE HAMBURGER APPROACH

We're starting with a prime cut - a very large number of respondents from five different specialties, only two of which are combat specialties. Since this is both a visable project and one in which the officers are personally interested, we expect a high return rate. At the time of this writing there has been a return of more than 70% of the first specialty data. The CONUS data has a greater than 80% return; it is expected that the overseas return rate will soon be similar.

Although there are other scales which ARI might have tried on an experimental basis, we will have two incumbent scales to compare for combat specialties. Relative Time Spent is given to only combat specialties, so we will not be able to compare it with non-combat specialties. One of the scales, Part of Position will be used for all specialties. It may be possible to compare across specialties using the "common" tasks.

Commonality across jobs is an important concept which ARI initially addressed with our Duty Module research for assignment purposes. Implications for training are obvious. With Army training diversified in many locations, recognition and management of the common areas can greatly reduce resource expenditures. TRADOC has appointed a proponent for each common task area. This proponent will perform the task analysis and develop the training module for each of its common tasks. All of the other schools will then use this "package" in their location. Although the common tasks were first determined to be common through the "smoke-filled room" method, the commonality will be verified by the survey.

Incorporating a common set of tasks in all the surveys enhances ARI's ability to evaluate some aspects of the survey. The original set of common tasks was so long there would have been no room for specialty tasks. In the process of attempting to refine the list to an acceptable number, the common tasks were rewritten to a very general level of specificity. The specialty tasks, however, are usually much more specific. Again, the total number of common plus specialty tasks was so large that many of the specialty tasks were rewritten to a more general level. We will have several levels of specificity to compare statistical reliabilities, even though there was no actual planning for this possibility.

Every officer in each specialty will be surveyed. This gives us a criterion measure for comparing other sample sizes. We can compute a definitive sample size which would have given data not significantly different from the total sample.

The final evaluation plan has these specific objectives:

a. Reliability of data: The most important objective is to determine whether job data gathered by the task item method produces reliable information. If reliabilities are not acceptably high, objectives b., c. and d. will not apply.

b. Survey length: The surveys contain 1200-1700 items to which each officer must respond. Reliabilities from objective a. will be examined to determine whether survey length affects reliability.

c. Specificity: Items and blocks of items which are written at different specificity levels will be compared to determine which produces the most reliable data. It is also possible that analysis will show that some items may be combined into one more general item.

d. Sample size: TRADOC presently plans to conduct a census of the Army officers. ARI analysis will determine the smallest sample size actually required to provide essentially the same information.

All of the desired objectives are not in this final set. However, the large sample sizes and differing types of specialties lead us to believe that we will have a very good grade of hamburger.

SUMMARY

All the surveys have not been returned from the field. Therefore, data analysis has not been started. We are not sure what to expect. If ARI were doing this project in a research mode, the filet mignon approach would have been used (including using other data collection methods), but the sample sizes would necessarily have been small. Also, the small number of available personnel would have forced a much longer time line.

This is the first time that the Army operational organizations and ARI have cooperated on a mutual program (to this extent). Although this project necessitated use of the hamburger approach, ARI has been included in the decision processes during the second year of the project. The synergestic effects of operation-oriented and research personnel should produce a quality product. We hope to be able to label it "choice grade."

APPENDIX A

PILOT OFFICER JOB SURVEYS

SCALES	11	13	31	73	214EV
	INFANTRY	FIELD	LAW	MISSLE	PERSHING
		ARTILLERY	ENFORCE	MATERIEL	REPAIR
				MNGMT	(WARRANT)
INCUMBENT					
Part of Position	2,000	1,500	460	190	60
Relative Time Spent	2,000	1,500	460	0	0
SUPERVISOR					
Training Emphasis	100	100	100	0	0
Task Learning Difficulty	100	100	0	0	0
Consequences of Inadequate Performance	100	100	0	0	0
Task Delay Tolerance	100	100	0	0	0
TOTAL RESPONDENTS	4,400	3,400	1,020	190	60

References

Cunningham, J. W. and Drewes, D. W. Determining the Training Requirements of United States Coast Guard Warrant and Commissioned Officer Billets. Proceedings of the 20th Annual Conference of the Military Testing Association, 1978, 28-50.

Gilbert, A. C. F.; Waldkoetter, R. O.; Raney, J. L.; and Hawkins, H. H.; Efficicy of a Training Priorities Model in an Army Environment (ADA066784); Technical Paper 343, US Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA 22333, October 1978.

Hemphill, J. K. Dimensions of executive positions; a study of the basic characteristics of the positions of 93 business executives. Research Bulletin 59-5, Princeton, NJ: Education Testing Service, 1960.

Mitchell, J. L. The Study of Executive and Managerial Jobs. The Study of Air Force Jobs: Symposium Papers. Lackland Air Force Base, Texas: USAF Occupational Measurement Center, 1978, 78-01, 17-22.

Mitchell, J. L. Differential Responses on Alternately Anchored Job Rating Scales. Proceedings of the 20th Annual Conference of the Military Testing Association, 1978, 525-536.

Ruck, H. W. The Collection and Prediction of Training Emphasis Ratings for Curriculum Developments Proceedings of the 20th Annual Conference of the Military Testing Association, 1978, 242-251.

Siebold, G. L. The Applicability of the ISD4-Factor Model of Job Analysis in Identifying Task Training Priority in Nine Technical MOS, US Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA 22333, submitted for publication.

Siebold, G. L., and Waldkoetter, R. O. The Function and Scope of Job Analysis in the Army, US Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA 22333, submitted for publication.

Van Nostrand, S. J. and Wallis, M. R. Occupational Analysis for Field Grade Army Officers. Proceedings of the 20th Annual Conference of the Military Testing Association, 1978, 373-384.

OFFICER JOB ANALYSIS:
THE ARMY'S OPERATIONAL PROGRAM

MAJ Grover A. Josey, Jr.

United States Army Training and Doctrine Command
Training Developments Institute
Fort Monroe, Virginia 23651

INTRODUCTION

To the Army, slicing the officer job analysis boloney is not just an academic question, to be discussed in conferences of occupational researchers but otherwise consigned to the "too tough" file. It is a question of immediate, practical concern.

The Army Training and Doctrine Command (TRADOC), in conjunction with other Army trainers and the Military Personnel Center (MILPERCEN), has begun a long range, phased program to analyze officer jobs in nearly one hundred commissioned and warrant officer specialties. This program will provide the foundation for implementing a number of profound changes in the officer training and education system recently directed by the Army Chief of Staff.

In August 1977, the Chief of Staff appointed a study group, the Review of Education and Training for Officers (RETO), to determine officer training and education requirements and recommend policies and programs for the officer corps of the 1980's and beyond. In the course of its study, RETO examined the education systems of the Air Force, Navy and Marine Corps, and those of other nations, industry and the professions. In order to determine training requirements for Army officers, RETO conducted an abbreviated analysis of each specialty, utilizing the duty module concept that had been developed and refined by the Army Research Institute (ARI) for use in describing and clustering officer jobs. The training and education proponent for each specialty (in most cases, a TRADOC service school) was asked to identify the principal jobs in the specialty, prepare a list of duty modules performed in those jobs, and recommend the appropriate setting (resident course, correspondence course, on the job training, unstructured self study, etc.) for training each duty module. RETO used these abbreviated analyses as a basis for developing a new training and education system for Army officers.

The final RETO report was published in June 1978.(1) Following a year of further study and refinement by the Department of Army staff, the decisions of the Chief of Staff were recently announced. A number of structural changes will be made in the officer training and education system:

- New pre-commissioning programs will be tested and pre-commissioning curricula will be standardized.

- Initial entry training (the Officer Basic Course) will be lengthened.

- A short company pre-company command course for captains will be instituted. Consideration will be given to entirely replacing the current six month Officer Advance Course with a series of short functional courses attended on an as-required basis.

- Military Qualification Standards (MQS) will be established for pre-commissioning and for each specialty at the lieutenant and captain level, to specify clearly what constitutes qualification, and serve as the base document for integrating the training and professional development efforts of the individual officer, his commander, and the Army school system.

- The Combined Arms Service Staff School (CAS^3) course will be established to train all new majors in staff procedures.

- Battalion and brigade pre-command courses will be instituted for all arms and services.

In addition to the decisions on the structure of the system, the Chief of Staff approved a long range program of job and task analysis to determine the specific training requirements of officer jobs in all specialties and grades, warrant officer through colonel. TRADOC has been developing this operational analysis program over the past two years.

DEVELOPMENT OF THE PROGRAM

To understand the program developed by TRADOC, one must first understand how the Army is organized to conduct job analysis. Unlike the other services, where a central organization performs the analysis and then passes data to users in the personnel management or training business, responsibility for job analysis in the Army is divided among a number of agencies, each with its own unique perspectives and requirements. On the personnel management side, MILPERCEN is responsible for analyzing officer jobs with a view to verifying or modifying the structure of officer specialties which comprise the Officer Personnel Management System (CPMS). On the training side, TRADOC Headquarters establishes job analysis methodology to be used within the command, but responsibility for actually doing the analysis is vested in the TRADOC service schools, each of which is the training and education proponent for one or more specialties. Other Army trainers, for example the Army Logistics Management Center (ALMC), the Academy of Health Sciences (AHS), and the Judge Advocate General (JAG)

School, perform analysis for their own specialties or subject areas. ARI conducts research on occupational analysis for both the training and personnel management communities.

Since a comprehensive analysis of officer specialties involves so many different organizations, it was necessary to adopt a methodology that was understandable and acceptable to all; to avoid, where possible, unnecessarily reinventing the wheel. Fortunately, during the time that the officer program was being planned, TRADOC was developing a new regulation and accompanying handbook on job and task analysis [2,3], based on the job inventory approach of the Interservice Procedures for Instructional System Development (IPISD) model, but incorporating lessons learned and procedures refined over the past several years. Although these procedures were based primarily on enlisted experience, and have been incompletely tested on officer jobs, there is an obvious parallel.

The enlisted model was modified, where necessary, to make it appropriate for use in officer analysis, and program guidance was published in TRADOC Circular 350-2, Officer Job/Task Analysis and Training Development.[4] The Army Occupational Survey Program (ASOP), operated by MILPERCEN, was selected as the principal means of data collection in order to provide a common data base from which to make training and personnel management decisions. This was considered critical because improvements in either the training system or the personnel management system can be negated if the two systems are not mutually supportive. AOSP has the additional benefit of providing relatively inexpensive, quantifiable data from job incumbents, which will give the Army a more supportable basis for defending new officer training programs to resource managers at Department of Defense, Office of Management and Budget, and Congress.

One of the most significant modifications made to the enlisted analysis model was in the handling of common tasks. Much of what Army officers do is common, regardless of specialty and, to some extent, grade. However, because the Army school system is decentralized, instruction in common areas is far from standard. A principal goal of the officer analysis program is to determine specifically what is common, and to standardize instruction (and conserve resources) by assigning a single school to develop instruction in a particular common area for use by all.

The use of a separate survey of common tasks administered to the entire officer corps was considered, but rejected because data on the importance of common tasks relative to specialty unique tasks is necessary in making decisions on task selection for training and the site and setting of such training. Instead, at the beginning of the program, proponents for nearly 50 common subject areas were designated and asked to contribute tasks from their subject areas which by doctrine or practice they believe to be common to the company grade and field grade levels. After refinement, these standard lists of common tasks will be included in surveys of each officer specialty. Other tasks, the commonality of which is less clear cut (designated shared tasks for the officer analysis program), will be identified and discussed among the proponents as the program proceeds, and included in the surveys of appropriate specialties. Job incumbents will provide the definitive answer of what is common and to whom.

THE PILOT PROGRAM

In the early stages of planning the operational analysis program, the decision was made to conduct a pilot program, involving several specialties, to verify/modify methodology and milestones before proceeding with full scale analysis. It was further decided that the pilot should be in two parts; Pilot I—company grade, and Pilot II—field grade. This was done because there is a more immediate need for analysis of company grade positions, and because the jobs of junior officers are more like enlisted jobs, and therefore presumably easier to analyze, than field grade positions.

Four commissioned officer specialties and one warrant officer specialty were selected for the company grade pilot, which began in October, 1978. They include large and small specialties, technical and nontechnical specialties, and at least one representative from each of the major branch groupings of Army officers; Combat Arms, Combat Support, and Combat Service Support. Table 1 provides a brief comparison of the Pilot I specialties.

For management purposes, the pilot was organized in eight phases:

Inventory Development
Questionnaire Development
Coordination
Final Technical Review
Survey Administration
Data Reduction
Data Analysis
Evaluation

Space limitations prohibit a full description of all of the phases, so only the key elements of each will be discussed.

Inventory development began with a review by the TRADOC Service Schools of authorization data to determine positions to be analyzed. This was followed by a comprehensive search of job and task information, including a review of doctrinal and technical literature and data collected by the RETO study, and interviews with panels of subject matter experts, master performers, and job incumbents and their supervisors. It quickly became apparent that officer jobs could not be adequately discribed in terms of task statements alone, so lists of skills, knowledges, responsibilities and equipment asso-

ciated with one or more tasks were also compiled. Although these additional lists are not comprehensive, they will provide the start point for the learning analysis which will follow job and task analysis.

After the inventories were completed, the TRADOC schools worked closely with MILPERCEN to develop officer questionnaires for each specialty. Table 2 shows the sections developed for each questionnaire. The draft questionnaires were then coordinated with a number of agencies in TRADOC and the Department of Army staff, and tested on a sample population of Officer Advance Course Students. Following final revisions, the questionnaires were printed and are now being administered to a census of company grade positions in each of the pilot specialties.

Two scales, relative time spent and part of position, are being tested on job incumbents. Four additional scales, training emphasis, task learning difficulty, task delay tolerance and consequencies of inadequate performance, are being tested on a small sample of supervisors. Table 3 lists the scales used for each specialty. Following the pilot, the best scale or combination of scales will be selected for the full operational program.

After the return of questionnaires from the field, and manual and optical scanning of the booklets, data will be compiled and displayed utilizing the CODAP computer programs and other software currently under development. TRADOC and MILPERCEN will then jointly and separately analyze the data. The joint analysis is designed to produce mutually supportive decisions on specialty structuring and training programs.

Since operational officer analysis on the scale being undertaken by the Army is unprecedented, a complete evaluation of the pilot program is crucial. The evaluation has, in fact, been underway since the pilot program began. In addition to evaluations by TRADOC and MILPERCEN, ARI is conducting research on the reliability of data and has been asked to assist in refining or modifying methodology, if appropriate. TRADOC is also sponsoring research on analysis of so called "soft skills" which will be integrated into the officer analysis program as the program becomes fully operational.

THE FUTURE

The symposium for which this paper was prepared posed the question, "How thin can the officer job analysis boloney be sliced". The TRADOC response to that question is, "As thin as we can get it". Pilot I will be completed in mid FY 80. Pilot II results will be in about 18 months later. The full operational program will extend through FY 85. Throughout this period we will be making modifications to both the methodology and milestones of the program. Although it is too early to report on results, preliminary indications are encouraging. We will keep you posted.

TABLE 1

OFFICER ANALYSIS PILOT SPECIALTIES

SPECIALTY	POSITION DENSITY	TECHNICAL CONTENT
11- Infantry	4150	Moderate
13- Field Artillery	3100	High
31- Law Enforcement	900	Low
73- Missile Materiel Management	170	High
214E- Missile Systems Technician, Pershing	78	High

TABLE 2

ORGANIZATION OF OFFICER SURVEY BOOKLETS

SECTION I	Background Information
SECTION II	Activities (Tasks)
SECTION III	Position Responsibilities and Requirements
SECTION IV	Additional Skills and Duties
SECTION V	Equipment
SECTION VI	Personal Comments

TABLE 3

OFFICER SURVEY SCALES

	OFFICER SPECIALTIES				
RATING SCALE	11	13	31	73	214E
Part of Position	X	X	X	X	X
Relative Time Spent	X	X	X		
Training Emphasis	X		X	X	
Task Learning Difficulty	X		X		
Task Delay Tolerance	X				
Consequences of Inadequate Performance	X				

REFERENCES

1. Report of the Review of Education and Training for Officers (RETO), Headquarters, Department of the Army, 30 Jun 78.

2. TRADOC Regulation 351-4, Job and Task Analysis, 9 Mar 79.

3. TRADOC Pamphlet 351-4 (Test), Job and Task Analysis Handbook, Aug 79.

4. TRADOC Circular 350-2, Officer Job/Task Analysis and Training Development, 1 Mar 79.

THE NEGLIGIBLE EFFECTS OF PROCESS-PRODUCT DISTINCTIONS ON EVALUATION QUALITY

John A. Boldovici
and
Eugene H. Drucker

Human Resources Research Organization (HumRRO)
Fort Knox, Kentucky 40121

Much has been written and said during the past few years about process and product measurement. The purposes of this paper are to identify sources of confusion about the process-product distinction, and more importantly, to show that the process-product distinction has little or no effect on evaluation quality.

The difference between process and product measurement obviously is not in the measure. One cannot tell whether 2.5 cm is a process or a product measure any more than one can tell whether 2.5 cm is a measure of the radius or the diameter of a circle. If not by inspection of a measure, how is the distinction to be made? It can be made along at least five dimensions. All writers are not in agreement as to which dimensions should be used (compare, for example, Osborn, 1973; with Army, 1977), and some writers switch from one dimension to another, without acknowledging the implications of having done so. (See, for example, Osborn, 1973; and Swezey and Pearlstein, 1975). The five dimensions along which the process-product distinction is made are shown in Table 1, and involve measuring:

1. By observation of behavior, as opposed to some trace ("product") of behavior. Guidance for developing Skill Qualification Tests (Army, 1977; Campbell, Ford, and Campbell, 1978) is unequivocal in distinguishing between process-scoring and product-scoring depending on whether,

    > ... the scorer observe[s] ... the task performance or ... some product produced during performance of the task (2, p. 4.7).

    The distinction seems reasonable and is easy to apply. Its adoption has not been widespread.

TABLE 1

FIVE WAYS TO DISTINCUISH BETWEEN
PROCESS AND PRODUCT MEASURES

Process If The Measure Is Of:	Product If The Measure Is Of:	References
Behavior	Trace ("Product") of Behavior	Army, 1977; Campbell, Ford, and Campbell, 1978
Performance of a Task Which Generates No Products	Performance of a Task Which Generates Products	Osborn, 1973; Swezey and Pearlstein, 1975
An Independent Variable or Treatment	A Dependent Variable or Outcome	Fitzpatrick, 1970
Use in Diagnosing Performance	Use in Certifying Performance	Osborn, 1973
Means	Ends	

2. Performance of tasks which do not generate "products" as opposed to tasks which do. Osborn (1973), and Swezey and Pearlstein (1975) referring to Osborn's work, distinguish between tasks in which the process is the product (close-order drill and diving, for example), and tasks in which the product always follows correct performance of the process (packing a parachute, for example). Osborn (1973) also notes, "Relatively few tasks are of the first type--those in which the product and the process are the same." Indeed, one might argue that no tasks are of the first type. Diving is no more the product of diving than parachute-packing is the product of parachute-packing. If a packed parachute is the product of packing a parachute, then a splash in the water or a wet diver must be the product of diving. The difference between the tasks is not that the product is the process in one and not in the other. Rather, one difference is in the concreteness or permanance of parachutes as compared to splashes--a difference with no compelling implications for measurement, because it is easily eliminated by the use of photography or other means of making permanent records. A more important difference is that one can, with a minimum of practice, inspect parachutes and make useful inferences about the quality of task performance, but cannot inspect splashes or wet divers and make very useful inferences about diving. This difference inheres, not in the tasks themselves, but in our ignorance about the relations between task outcomes and performance quality. Where little or no special knowledge is required to relate outcomes to performance, the performances typically are judged on the basis of outcomes. Footraces are examples. Where special knowledge is required (and absent), we relegate performance assessment to judges and critics. Evaluating diving, concerts, and gymnastics are examples.

3. Independent, as opposed to dependent, variables. Fitzpatrick (1970), in discussing measurement for program evaluation, notes that process measures describe the program (the independent variable) as it occurred, and outcome measures (the dependent variables) identify program effects.

4. For diagnosis, as opposed to certification. Osborn (1973), in describing the uses of process and product measurement, introduces a distinction based on test purpose:

   > ... measures which focus on task outcomes (products) normally provide data relevant to the first purpose [certification], whereas measures of how tasks are carried out (process) pertain to the second [diagnosing instructional weaknesses] (p. 1).

   This is a variant of another dimension for making the distinction; namely, a means-end dimension.

5. Means, as opposed to ends. Since processes and products are roughly analogous to means and ends, viewing all measures of means as process measurement, and all measures of ends as product measurement seems reasonable. The means-end dimension is useful because it implies the need to specify objectives: without objectives, one cannot determine whether means or ends have been measured.

Distinguishing between process and product measurement is not difficult if the measured event or product falls at the same end of all five dimensions. If, for example, a measure is made by direct observation of performance of a task which has products from which inferences cannot be made about performance quality, and the measure is of an independent variable and used diagnostically, and the task is a means to an end, then one is clearly dealing with a "process" measure. Confusion arises, however, when a measured event or product falls near opposite ends of any two dimensions. Suppose for example that measures taken from a sight photograph were used diagnostically. The SQT experts would say that product measurement had taken place, because observation was not made of behavior. The measures would, however, qualify as process if the diagnosis-qualification dimension were used. Whether the measures were of means or ends would depend on the stated objective: means (process) if the objective were to hit a target, ends (product) if to lay cross-hairs on an aiming point.

Despite apparent confusion about appropriate dimensions for distinguishing between process and product measurement, the distinction may be useful--albeit indirectly. The distinction reflects a concern with objective, as opposed to subjective evaluation. If evaluation can be defined as making judgments based on comparisons between expected or desired characteristics of behavior or a product of interest, and observed characteristics of the same behavior or product, then it seems to follow that the more precisely stated the expectations and the more objective the observations, the better the evaluation. This line of thinking underlies current emphases in criterion-referenced testing: stating standards (expectations) on the one hand, and objectively measuring performance (or "products" of the performance) on the other. Notice, however, that standards can be well defined, loosely stated, or not defined or stated at all. Examples near the extremes are, "At least one target hit with two rounds," and "neutralizes targets at various ranges." Similarly, behavior or product characteristics can be measured precisely, measured imprecisely, or measured not at all. Examples near the extremes are measuring length with a ruler and estimating casualties. Evaluation improves with the extent to which it consists of comparing measured characteristics of behavior or products with well defined standards.

Figures 1 and 2 present examples of kinds of evaluations which result from various combinations of precision and imprecision in standards on the one hand, and measurement or estimation of behavior or products on the other. The upper left cell in both figures is the most desirable from the standpoint of good evaluation. Here measured characteristics are compared with well defined standards. When a physician says, "Your pulse rate is high," we have confidence in this evaluation because (a) the physician has measured our pulse rate objectively, (b) the standard is clearly defined, and (c) the rules for making a judgment of high or low are obvious. For the same reasons we have confidence in evaluations of tensile strength (Figure 2, upper left cell.)

The upper right cells in both figures inspire less confidence, but retain the virtue of well defined standards. Estimates rather than measures of behavior or product characteristics are used here, probably for expediency, and could, with added cost, be replaced by measurement: A photograph could be made of the pass receiver and the boundary line (Figure 1), and instruments could be used to measure front-end play in the car (Figure 2). Comparing the results of the photographs or of the instrument readings with available standards for "out-of-boundness" or the need for front-end work would increase our confidence in the evaluation.

The bottom cells of the two figures are of interest mainly as areas to be avoided by evaluators. In the lower left cells the evaluator makes precise measures of the behavior or product of interest, and compares these with a standard which is loosely or not defined. "Evaluations" of mental health and of stereo equipment, for example, involve very precise measurement of many behavior and product characteristics whose relevance to a behavior or product standard cannot be known because the standard itself is not known, and the rules for making the comparisons are nearly as numerous as the evaluators.

		BEHAVIOR CHARACTERISTICS	
		Measured	Estimated
BEHAVIOR STANDARD	Well defined	Comparing observed with normal heart rate.	Referee calling a pass receiver out of bounds.
	Loosely or not defined	Evaluating a patient's mental health based on results of diagnostic tests.	Judging ice-skating, diving, dancing.

Figure 1. Four kinds of process evaluation.

		PRODUCT CHARACTERISTICS	
		Measured	Estimated
PRODUCT STANDARD	Well defined	Comparing measured tensile strength with design specifications.	Deciding whether a car needs front end work by shaking its wheels.
	Loosely or not defined	Consumer-oriented tests of stereo equipment, TV, washing machines, etc.	Judging a painting or a novel.

Figure 2. Four kinds of product evaluation.

In the last (bottom right) cells of both figures the evaluator is comparing the results of no measurement with a nonexistent or ill-defined standard. This is the realm of opinion and sophistry.

Perhaps the most interesting aspect of Figures 1 and 2, however, is that good evaluation is not the result of whether process (behavior) or product (trace of behavior) has been measured. It is rather a function of:

1. Precision in stating standards.

2. Objectivity in measuring the product or behavior characteristics.

3. Veridicality of rules for comparing measured characteristics and standards.

This is so regardless of whether processes or products are being evaluated.

HANDS ON TESTING: A MODEL FOR SCORER TRAINING

Roy C. Campbell

Human Resources Research Organization (HumRRO)
Radcliff, Kentucky 40160

INTRODUCTION

During the mid 1970s as performance testing emerged as the primary means of soldier evaluation, the emphasis within the military concentrated primarily on the role of the developer in the system. Many guidelines and training courses were prepared to assure these individuals delivered a quality product. As a result there are many competent test developers in service schools, training headquarters and test development agencies. These developers spend long hours producing tests that are accurate measures of performance. More time is spent on validating the test instruments in formal and informal tryouts using representative incumbents and trained test administrators or scorers. By the time a test developer completes the development cycle the end product is a finely tuned instrument that has usually gone through many revisions to eliminate any source of scorer unreliability and measurement errors. The test developer has done his best and his test is ready to be fielded. Yet there is one characteristic of most large scale military testing that can work to negate the efforts of the developer. That characteristic is that military testing is almost always designed to be decentralized. The product of the test developer is turned over to others--usually in the examinee's unit--for administration and scoring. All the reliability that the test developer designed into the test can be lost if that scorer does not administer and score the test in the way the test developer envisioned. No matter how competent scorers are in other aspects, they cannot administer and score a performance test reliably--even the simplest performance test--without some training. What follows is a description of one model for scorer training that was designed to enhance scorer competence and hence test reliability.

Since 1975, the Human Resources Research Organization (HumRRO) has been working with the Individual Training Evaluation Directorate of the U.S. Training and Doctrine Command (TRADOC) through contracts administered by the Army Research Institute in various aspects of development and implementation of the Army's Skill Qualification Test (SQT). This scorer training model was developed as one aspect of that work.

ꞮHE SCORER TRAINING PACKAGE

The Army's SQT system is a complex, decentralized testing system designed to evaluate its enlisted population on job skills. Approximately one third of any individual test is in hands on performance. Very early in development work on SQT it became apparent that if these hands on tests were to be administered reliably at the unit test site some means must be used to insure that the test scorer was operating on the same "wave length" as the test developer. Here, a word about the mechanics of the administration of SQT is in order. In the SQT system a Test Control Officer (TCO) has responsibility for administration of all SQT either by units or over a geographical area. Working with individual unit commanders, the TCO insures the appointment of a Test Site Manager (TSM) who has the responsibility for administration of a specific test. Under the control of the TSM are scorers who have responsibility for the administration and scoring of a specific task. For each task there is also the minimum of one alternate scorer (see Figure 1). The majority of these scorers are in the pay grade E-6 and usually of the same military occupational specialty (MOS) as those being tested. This then is the framework for SQT Administration.

As it became apparent that untrained scorers could not score tests reliably, it also soon became readily apparent that the training could not be left to unit discretion if any type of thoroughness and uniformity was to be ahcieved. While the TSM would be the conduit for the training, only the test developer knew enough about the variations and critical points of the test and it must be his responsibility to develop and organize the training that the TSM would deliver.

The training package for preparing scorers to operate in this framework had to meet three criteria:

1. It had to be short--not exceeding 4 hours per SQT.

2. It had to be meaningful; that is, it had to relate directly to the test being administered.

3. It had to allow flexibility in administration and evaluation.

The training package developed was divided into two parts. The first part consists of a briefing by the TSM during which the test station assignments for the scorers are made. This is followed by a standardized training VTR which demonstrates the general principles of scoring. Scorers are provided guidance checklists which they use to evaluate the filmed situations during a TSM led critique.

The majority of the training period is spent in a practical exercise phase requiring approximately 3 hours of time. In order to enhance learning and maximize retention of skills it is stressed that the practical exercises should be held within 24 hours of the actual SQT. In practice, the training session is easiest to conduct just before the formal tests, since it requires the same equipment and test setup as the test. The training session then serves as a dress rehearsal or dry run for the unit administering the SQT as well as for training scorers.

The time requirement is met by matching scorers and alternate scorers for each test station into a training group. If the size of the SQT dictates multiple replications of a test setup, all like scorers are grouped together. All scorers (including alternates) for the station train together with one person serving as a mock examinee and the other scorers observing and evaluating the scoring. If there are no additional personnel the TSM and his assistants (if any) rotate between the test stations as evaluators and the mock examinee also provides a more informal critique of the scorer's performance. Scorers rotate through the positions of primary scorer, mock examinee and evaluator. Since the training is designed into a self contained scorer training package, all stations in an SQT can train simultaneously. The TSM should observe each scorer and alternate scorer before certifying the individual as ready to perform as a scorer.

Meeting the second goal--making the training relevant--is assured by requiring the test developer to adhere to some strict guidance in the development of the scorer training package. The scorer training package is to be provided to the TSM as Chapter 8 of the Manual for Administration of the Hands-On Component which is the formal document that covers all aspects of SQT hands on testing, scoring and administration (see Sample Scorer Training Package). The purpose of the scorer training package is to guide those individuals who are role playing the part of the "examinee" during the training. This is done by developing errors or actions that the mock examinee must commit or perform during the test. Since each scorer is critiqued on his performance--usually by the mock examinee or the TSM--developers also list the action the scorer should have taken for each error to give the evaluator a basis for the critique.

Care and consideration are required when a developer selects the actions of the mock examinee. The actions must be realistic and the role playing must not be too contrived or sophisticated. Because the scorers and alternate scorers are usually contemporaries, the role playing can easily become a parody that loses the sense of realism. Besides being practical, examinee actions must have training value. Some considerations for selection of errors or actions are:

- The most likely errors that examinees will commit should be selected. The scorer must be alert to these because of their frequency even if they are not difficult to detect.

- Actions that cause problems during development should be errors in scorer training. During the construction of test items and the corresponding scorer instructions the developer usually will have experienced problems in one or more areas. These problems could vary from difficulty in wording of the performance measure to variations in outcome if test conditions are not standardized. Areas that cause problems during development are likely areas of problems during the test. These should be highlighted as mock examinee actions.

. Scorer training should reinforce the scorer's instructions instead of merely reflecting the opposite of the performance measure. If a performance measure has detailed corresponding scorer's instructions, it should receive more emphasis in scorer training than performance measures that are clear-cut and require no scorer's instructions. Variations covered in the scorer's instructions to meet different situations should be included in the scorer training, especially if there are a variety of correct ways to do a step.

. Unusual occurrences other than errors should be included. Examinee actions need not be limited just to the performance measures. Some other actions that could be included are:

- Ask questions after the Examinee Instructions are read or during the performance.

- Ask for assistance from the scorer or for unauthorized assistance from an assistant such as a driver.

- Call a malfunction in the equipment when none exists.

- Introduce, if possible, a real malfunction into the equipment.

- Perform some step that is outside the scope of the test.

- Challenge the evaluation (rating) of the scorer even when an "error" is committed.

Some tasks are not so difficult that a comprehensive list of errors or actions can be developed. But there should be at least two recommended actions per task to allow a variety of performance during the practice. If no actions stand out as potential problems, the errors should cover the range of the task to insure the sample covers the different aspects of the task.

Since many steps in a task have an impact on subsequent actions, the entire task must be considered in view of the introduced action. The action or error must be sensible. For example, some developers have approached their selection by instructing the mock examinee not to perform one of the initial steps. When this is done, the task stops because the examinee can go no farther. This is not a very realistic occurrence.

It is better to direct the examinee to perform something incorrectly, if possible, than to merely omit a step. Likewise, the action directed must be specific. It is inadequate to state in the instructions to the mock examinee that he should, for example, "Perform the disassembly incorrectly." The instructions must specifically state what the mock examinee is to do or not to do. Likewise, the instructions to the mock examinee must specify exactly where the occurrence is to take place. For example, "Do not place the weapon on SAFE when clearing prior to disassembly."

Finally, developers are cautioned that they must be wary when introducing actions or errors that could damage equipment or be a potential cause of personal injury. This is sometimes a difficult decision because in some tasks very realistic actions on the part of examinees present these problems. So the developer must weigh the use of such actions against the likelihood of their occurring during the test.

The scorer training session is intended to be as freeflowing and realistic an exercise as possible. It is not supposed to be a canned exercise with the scorer knowing what errors are to be committed when. The developer provides the examinee with a list of actions which he can choose from, or he can choose to perform the entire task correctly, or he can choose to commit an action not on the list. For this reason, it is not necessary to break out the errors to be committed by the scorer or alternate scorer who is role playing the examinee. Even though the scorer may have access to the scorer training guidance sheets, the aim is still to present unexpected occurrences.

Developers must also prepare a scorer action column that states what the scorer should do for each action the examinee commits. This is used for critique purposes. In many cases, the scorer action is very straightforward, such as, "Mark performance measure Fail." But in other cases, for example, when the mock examinee action is not wrong but appears wrong, the Scorer Action explanation must be more detailed. For example, the mock examinee may be directed to perform part of the task out of sequence when sequence is not a scored requirement. The proper scorer action in that case would be to mark the performance measures Pass with the explanatory note that sequence is not a scored behavior.

The final criterion for scorer training--that of flexibility--is met by allowing the TSM to adapt the package to his own needs and more specifically to the needs of his scorers. The scorer training package ideally provides all the tools the TSM needs but he is free to use those tools as he sees necessary. For example, not all SQT tasks are equal. Many present no unusual scoring problems. In those cases the TSM can have the scorer and alternate scorer perform the minimum training which is that each perform the task once and each score the task once. As scoring difficulty increases the number of repetitions of the scoring practice, each with a variation of a typical problem can be increased. Likewise, the use of the package can be adjusted to meet the variation in tasks. For some tasks, the TSM can give the entire package to the scorer/alternate scorer and they can self-administer the training. Although this means that the scorer can read in advance what errors or actions the mock examinee can commit, the exact action is not known and the mock examinee does have the option of performing the task correctly. For other tasks the TSM may retain the part of the package containing the examinee actions and verbally instruct the mock examinee on what action to commit. Finally, the TSM has the option within the framework of the package of introducing mock examinee actions which are not listed by the developer which may be more appropriate to the TSM's particular situation.

The last way in which the TSM can adapt the scorer training package is through the use of the package as a certification tool. Certification of individuals as ready to perform as SQT scorers is a TSM responsibility and perogative. There are currently no outside criteria of what a scorer must do for certification. The TSM is thus free to apply the scorer training package in this role as he perceives the needs and capabilities of the individuals he has been given as scorers. Since the TSM must be ultimately held responsible for the conduct of the actual test he should be free to determine the competency of the scorers before the test. The tools for this determination exist in the scorer training package.

CONCLUSIONS

The advantages of the scorer training package should be obvious. It presents a systematic reasoned duplication of actual testing occurrences in a training environment. It also assures scorer familiarity with the task being tested, a condition that is often assumed by TSM but unfortunately not always warranted. Finally, it allows the TSM the opportunity to pre-evaluate his scorers and replace some if necessary or, short of that, to alert him to stations that may need extra attention during the actual test.

The scorer training package has two disadvantages. First, even though it is coupled with other necessary preparation for the SQT--namely the setting up of the test station--it still requires some extra time on the part of the TSM and scorers. And time is always a premium commodity in any unit. Second, the very flexibility that is encouraged in the application of the scorer training package may tempt the TSM to bypass it completely. We have seen this occur during some of the validation tryouts of the tests and the results have reinforced our beliefs in the need for the training. The use of the scorer training package can only be maximized as commanders and TSM become aware of its value as a link in the chain of scorer reliability and efficient, accurate test administration.

SQT Administration Chain*

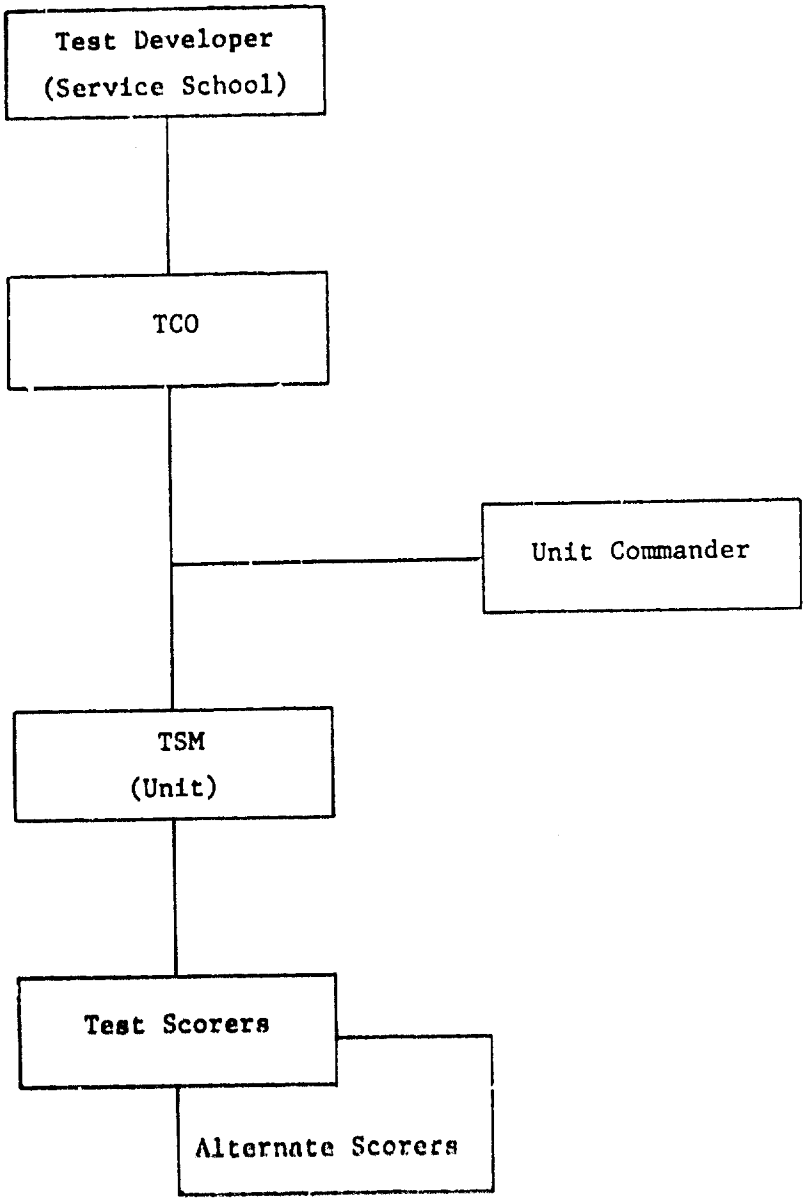


Figure 1

(*Diagram is schematic only and does not reflect the actual relationship between individuals.)

STATION 1

GUIDANCE SHEET FOR SCORER TRAINING: MAINTAIN A COAX MACHINEGUN

You are either the primary scorer or the alternate scorer for this station. During this training session you will serve as both the scorer and the "examinee." When you are the "examinee," you will also be the evaluator for another person who is the scorer as you perform the task.

SCORER: When you are the scorer during this training session, you must perform the following:

a. Set up the test station as the scorer's instructions direct. Conduct this test exactly as you will when you administer it to others for record.

b. Administer the test as the scorer's instructions direct. Even though the "examinee" will be working with you throughout this training session, treat him like he was a real examinee. You are being evaluated on your performance.

c. Score the "examinee's" performance by closely observing his actions and following your scorer's instructions. The "examinee" may perform a variety of actions that may or may not be correct. The purpose is not to trick you but to prepare you for what may occur when you administer the test for record.

d. After the administration of the test is over, you will be critiqued. Discuss with your "examinee" and other evaluators what he did and why you reacted as you did. If you do not feel comfortable with your role as scorer, ask to have the test repeated as many times as necessary. By the time this training session is over, you should know how you plan to administer this test for record and how you will react when the unexpected occurs.

EXAMINEE: When you are the "examinee" during this training session, you must perform the following:

a. Read the guidance sheet on the next page. On it is a list of "Examinee" Actions. During the test you may:

   (1) Commit one or more of the actions listed on the Guidance Sheet, or

   (2) Perform the task correctly, or

   (3) Commit an error that is not listed. If you do this, do not do anything to damage equipment. Also, be sure the error is realistic. Remember, you must be able to tell the scorer if his reaction to your error was correct.

b. Review your scoresheet. If the scorer marked you FAIL on a step you thought you did correctly, find out why.

c. Critique the scorer after the test is over. Across from the "Examinee" Actions are the actions the scorer should have performed. Use these to guide your critique for specific actions.

d. Discuss the scorer's performance in general. Make recommendations that you observed that will make his performance better. He will be doing the same for you.

e. Repeat the test several times. Vary the errors you commit each time you perform the test.

GUIDANCE SHEET FOR STATION 1

MAINTAIN A COAX MACHINEGUN

"EXAMINEE" ACTION	SCORER ACTION
1. After the instructions are read, ask the scorer if you must inspect the weapon.	1. Rereads the instructions.
2. When clearing the weapon, do not place the safety in the "SAFE" position.	2. Marks PM 1c NO. Marks PM 1 FAIL.
3. When clearing the weapon, look into the chamber but do not feel it.	3. Marks PM 1f NO. Marks PM 1 FAIL.
4. During disassembly, do not remove the barrel from the barrel jacket.	4. Marks PM 2a NO. Marks PM 2 FAIL. Writes DID NOT REMOVE BARREL FROM JACKET in REASON section.
5. After you start disassembly, ask the scorer if you must remove the charger assembly. (Do not remove it regardless of what scorer says.)	5. States, "I AM NOT ALLOWED TO HELP YOU IN ANY WAY. CONTINUE WITH THE TEST."
6. When performing the function check, after charging the weapon, place the bolt forward. Then charge the weapon, place the safety on SAFE and attempt to fire with the manual trigger.	6. Marks PM 4i NO. Marks PM 4 FAIL.
7. After performing the function check, repeat the procedure for clearing the weapon.	7. None required. The examinee's action is not a cause for FAIL.

THE INSTRUCTIONAL QUALITY INVENTORY AS A FRAMEWORK FOR TESTING

John A. Ellis and Wallace H. Wulfeck, II.
Navy Personnel Research and Development Center

Introduction

Modern military instruction is developed according to systematic methods which include the following steps:

1. Job/task analysis leading to specification of instructional objectives;
2. Development of tests to measure student progress toward the objectives, or to diagnose partial progress toward the objectives;
3. Design of new instruction and/or adaption of existing instruction to achieve the objectives;
4. Implementation of the instructional program;
5. Evaluation and feedback for course maintenance.

All systematic instructional design methods attempt to insure that instruction is job-relevant, by deriving objectives from on-the-job performance, and by designing tests and instruction which are deliberately matched to the objectives.

It is obvious that an important part of these systems is testing. Tests and test items must be designed so that they (at least) accurately measure the behaviors specified in the objectives. (In addition, tests may also attempt to diagnose reasons for inadequate performance on specific objectives.) Since each objective is a statement of a "criterion" for student performance, tests for objectives are called "criterion-referenced" tests. A test item is "referenced" to an objective (criterion), when it is consistent with the objective. Consistency means that the conditions and standards in the objective are maintained in the testing situation, and that the same student behavior is required in the test and the objective. The problem to be addressed in this paper is, "How can criterion-referenced tests be constructed?"

It is important to realize that traditional psychometric approaches to test development have little to contribute to the development of criterion-referenced tests, primarily because they were not intended for this purpose. Most psychometric methods were developed to determine whether tests could discriminate between individuals; they were not concerned with how items are initially developed. Also, for criterion-referenced purposes, no statistical method can be used to select items on the basis of consistency with objectives, because such methods are insensitive to the conditions, standards, or behaviors required by instructional objectives.

The intent of this paper is to propose instead a logically (and in some cases empirically) based set of procedures or guidelines that should be followed to develop test items which are criterion-referenced. As stated above, the central problem is to develop items which "match" specific objectivas. The standard approach to this problem (e.g., Briggs, 1977; Gagne, 1976) involves classifying objectives according to some scheme or taxonomy, then developing items for objectives on the basis of the classification. For this approach to be successful, two conditions must be met. to classify objectives reliably. Second, the classification scheme must have clear implications for test item development; the implications should be specified as prescriptions for item development. Most existing schemes fail to meet either one or both of these conditions. Typically the classification procedures are too loosely defined (e. g. Gagne, 1976; Popham, 1977) to permit reliable classification, and/or the implications for testing are not clearly specified (Briggs 1977). In the following sections a classification scheme which hopefully meets these conditions will be presented and specific prescriptions for item development will be discussed. This scheme and the resulting prescriptions represent an idealized framework for test and test item development. It is recognized that there are constraints present in the "real world" which frequently prevent the test writer from achieving the ideal. Therefore, the final section of this paper will present methods for dealing with "real world" restrictions.

THE CLASSIFICATION SYSTEM:

The following classification system is applied to the three main parts of instruction: objectives, tests, and instructional presentations. Each objective, test item, or piece of presentation, can be classified according to:

1. What the student must do, i.e., the TASK to be performed, and
2. The type of information the student must learn, i.e., the instructional CONTENT.

These two classification dimensions can be combined to form the TASK/CONTENT MATRIX.

The TASK Dimension:

There are two main TASKS a student can perform:

1. He can REMEMBER information, or
2. He can USE the information to do something.

EXAMPLE:

Here are two test items:

1. The symbol for resistor is ____________.
2. Using your knowledge of electronic theory, what would happen in the circuit shown below if the load resistance were shorted?

These two test items differ with respect to what the student is supposed to do (TASK). In number 1, the student has to REMEMBER something, and in number 2, the student has to apply or USE his knowledge in a new situation.

The CONTENT Dimension:

There are five types of CONTENT:

FACTS are simple associations between names, objects, symbols, locations, etc.
CATEGORIES are classifications defined by certain specified characteristics.
PROCEDURES consist of ordered sequences of steps or operations performed on a single object or in a specific situation.
RULES also consist of ordered sequences of operations, but can be performed on a variety of objects or in a variety of situations.
PRINCIPLES involve explanations or predictions of why things happen in the world. That is, they concern predictions or interpretations based on theoretical or cause-effect relationships.

NOTE: Facts can only be remembered. The others can be remembered or used.

EXAMPLES:

The following examples illustrate the five content areas for the REMEMBER task level:

REMEMBER FACT
1. The symbol for resistor is _______.
2. The student will list the names of the parts in the wind indicating instrument.

REMEMBER CATEGORY
1. List the defining characteristics of a jet pump.
2. The student will define the various kinds of clouds (cumulus, stratus, etc).

REMEMBER PROCEDURE
1. List in order the steps for cleaning an M-16 rifle.
2. The student will describe the procedure for preparing and sending a radio message.

REMEMBER RULE
1. List the steps involved in finding the rhumb-line course between two points on the earth.
2. The student will state the general rule for solving for circuit current, given voltage and resistance.

REMEMBER PRINCIPLE
1. State the principles of electron movement in a semiconductor junction.
2. The student will recall the reasons why hydraulic fluid contamination must be avoided.

Facts can only be remembered, but for the other content types, the student may be asked to USE his knowledge to classify, perform, solve, or predict. The following are examples of the USE task level for all content types except facts:

USE CATEGORY
1. Which of the pumps aboard ship are jet pumps?
2. Given photographs of clouds, the student will sort them according to type (cumulus, stratus, etc.).

USE PROCEDURE
1. Clean an M-16 rifle.
2. The student will prepare and send a radio message.

USE RULE
1. Calculate the rhumb-line course from Pearl Harbor to Long Beach.
2. Given the values for voltage and resistance, the student will calculate the current flow.

USE PRINCIPLE
1. Describe the theoretical movement of electrons in a PNP transistor.
2. The student will predict what is likely to occur if the landing gear fluid were contaminated.

The USE level can be further divided into two types:
1. USE-UNAIDED in which the student has no aids except his own memory.
2. USE-AIDED in which the student has a job aid for performing the task.

For this level, the nature of the aid depends on the content type:

For USE-AIDED CATEGORY the aid should consist of a decision strategy, including each critical characteristic, and the decision to be made according to presence or absence of that characteristic. In simple cases, the aid may only include a list of characteristics; the decision strategy is then implied.
For USE-AIDED PROCEDURES the aid would be a list of steps to be performed.
For USE-AIDED RULES the aid would be at least a statement of the formula or rule to be applied, and could include guidelines for when and how to apply it.
For USE-AIDED PRINCIPLES the aid would also be at least a statement of the principle, and could include guidelines for when and how to apply it.

EXAMPLES:

USE-AIDED: A pilot's preflight checklist is a USE-AIDED procedure. The pilot does not have to remember the steps or their order because they are on the checklist. The pilot does need to perform the steps correctly.

USE-UNAIDED: "The student will field-strip an M-16 rifle." Here, the student must remember the steps in the correct order, and perform them correctly.

In summary, the REMEMBER level involves "pure" remembering,
the USE-UNAIDED level involves remembering what is to be used, and then using it, and
the USE-AIDED level involves "pure" using.

The entire TASK/CONTENT MATRIX is shown below:

	FACT	CATEGORY	PROCEDURE	RULE	PRINCIPLE
REMEMBER	RECALL OR RECOGNIZE NAMES, PARTS, DATES, PLACES, VOCABULARY DEFINITIONS, ETC.	REMEMBER THE CHARACTERISTICS OF EACH CATEGORY AND THE GUIDELINES FOR CLASSIFICATION.	REMEMBER THE STEPS OF THE PROCEDURE.	REMEMBER THE FORMULA OR THE STEPS OF THE RULE.	REMEMBER THE CAUSE AND EFFECT RELATIONSHIPS OR THE STATEMENT OF THE PRINCIPLE.
USE UNAIDED		CLASSIFY OR CATEGORIZE OBJECTS, EVENTS, IDEAS, ACCORDING TO THEIR CHARACTERISTICS, WITH NO MEMORY AID.	APPLY THE STEPS OF THE PROCEDURE IN A SINGLE SITUATION OR ON A SINGLE PIECE OF EQUIPMENT, WITH NO MEMORY AID.	APPLY THE FORMULA OR RULE TO A VARIETY OF PROBLEMS OR SITUATIONS, WITH NO MEMORY AID.	USE THE PRINCIPLE TO INTERPRET OR PREDICT WHY OR HOW THINGS HAPPENED OR WILL HAPPEN, WITH NO MEMORY AID.
USE AIDED		GIVEN CATEGORY CHARACTERISTICS AND GUIDELINES, CATEGORIZE OBJECTS, EVENTS, IDEAS, ACCORDING TO CHARACTERISTICS.	GIVEN STEPS OF THE PROCEDURE, APPLY THE PROCEDURE IN A SINGLE SITUATION, OR ON A SINGLE PIECE OF EQUIPMENT.	GIVEN THE FORMULA OR RULE STEPS, APPLY THE FORMULA OR RULE TO A VARIETY OF PROBLEMS OR SITUATIONS.	GIVEN A STATEMENT OF THE PRINCIPLE, INTERPRET OR PREDICT WHY OR HOW THINGS HAPPENED OR WILL HAPPEN.

Any objective, test item, or piece of instruction will be classifiable in one and only one cell of the matrix above.

Important Implications of the Classification Scheme.

Remember the Job. The most important thing to remember when attempting to use the classification scheme is the job the student is being trained to do. The classification scheme was designed so that classification depends on the job requirements. The most important requirement to consider is whether or not the student will have to deal with objects or situations he has not seen or encountered during training. For the FACT and PROCEDURE content types, this does not occur. Facts by definition must be presented during training. The job requirements for procedures involve single pieces of equipment or single situations, and the student does not have to "generalize" to new equipments or situations. In other words, everything the student needs to know is presented during training.

On the other hand, there are some job situations that require the student to deal with so many possible objects, events, ideas, problems, or situations, that it would be impossible to include all of them during training. In this case, the training program is designed so that the student will be able to deal with new cases. CATEGORIES, RULES, and PRINCIPLES are used in the classification scheme to cover this situation.

The CATEGORY content type is used when the job requires that a large number of possible objects, events, etc. be classified into, or identified as a member of, one of a small number of particular categories. Instead of having to remember each object and its classification, the student is given characteristics for each category, which allow him to classify objects, etc., he has not seen before.

The RULE content type is used when the job requires that a large number of problems be solved or that a complicated sequence of steps be performed on a large number of different objects, events, etc. Instead of having to remember each problem or go through the steps on each object, the student is taught a RULE which allows him to deal with problems, objects, and events he has not seen before.

The PRINCIPLE content type is used when the job requires prediction or interpretation of a large number of possible situations, events, effects, etc. Instead of having to remember each possible situation or event and its effects, the student is given a PRINCIPLE which summarizes the "how" or "why" of general situations or which allows the student to predict what is likely to occur in a variety of situations.

TEST DEVELOPMENT AND THE TASK/CONTENT MATRIX

The classification scheme can be used to develop test items that are consistent with their associated objectives. This is accomplished by constructing each test item according to the steps listed below.

STEP 1. Write each item so that ACTION matches the action of the objective
Step 1a. Write each item so that the task level and content type of the items match the task level and content type of the objective.

STEP 2. Write each test item so that the CONDITIONS in each item, or the CONDITIONS under which the item is administered, match the conditions in the objective.

STEP 3. Write each item so that the STANDARDS in each item, and the STANDARDS for scoring each item, match the standards in the objective.

STEP 4. Write each item so that the format is appropriate for the task level and content type. Use the table below:

CONTENT TYPE

TASK LEVEL	FACT	CATEGORY	PROCEDURE	RULE	PRINCIPLE
REMEMBER	*for RECOGNITION:* matching true-false multiple choice *for RECALL:* short answer fill-in listing	short answer fill-in listing	short answer fill-in listing	short answer fill-in listing	short answer fill-in listing
USE-UNAIDED		performance matching true-false multiple choice short answer fill-in	performance true-false multiple choice short answer fill-in	performance true-false multiple choice short answer fill-in	performance true-false multiple choice short answer fill-in
USE-AIDED		performance matching true-false multiple choice short answer fill-in	performance true-false multiple choice short answer fill-in	performance true-false multiple choice short answer fill-in	performance true-false multiple choice short answer fill-in

EXPLANATION FOR THE CONSISTENCY PROCEDURE

STEP 1: The task/content level of the test item should match the task/content level of the objective. This means that the action verb in the test item should be the same as the action verb in the objective, or at least the same behavior must be required. If it isn't, the test item is measuring something different than was required in the objective.

STEP 2: The conditions in the test item, or the conditions under which the item is administered, should match the conditions specified in the objective. Naturally, there are some situations when, for reasons of safety or practicality or cost, testing conditions or on the job. In these cases, it is important to simulate the conditions as closely as possible. It is important to REMEMBER THE JOB; that is, the testing situation must be close enough to the job situation or later training situation, so that you can be sure that the student has achieved the objectives.

STEP 3: The standards in the test item, or the standards for scoring the test item, must match the standards in the objective. In criterion-referenced testing, standards are not arbitrarily selected. It makes no sense, for example, to require a student to get 80% of the items right, if he needs to recall all the information. On the other hand, for some tasks, a 70% or 80% criterion may be reasonable. In all cases, though, the standard specified in the objective should be used.

STEP 4: There are a number of different test item formats, and these may be more or less appropriate depending on the task/content level of the objective. The chart on the bottom of the previous page shows the acceptable formats for each task/content level.

Item formats at the REMEMBER level

In the chart for step 4, notice that at the REMEMBER level, recognition items (multiple-choice, matching, true-false) are usually not appropriate. This is because they don't test recall, only recognition. Most REMEMBER level objectives require recall because of the nature of the job.

Item formats at the USE level

Multiple-choice, matching, and true-false items can be appropriate, if carefully designed, for many USE-level tasks. For example, a category classification is often a true-false judgment. If the student must solve a math problem (Use-Rule), a multiple-choice item in which all alternatives are reasonable is appropriate. Also, some Use-Principle predictions involve a limited set of possible alternatives; again, multiple-choice is appropriate.

Why is format important?

The reason why test item format is important is that students are not dumb! The first thing most new students do in a course is find out how they will be tested. Then, they study just enough to pass the tests. If the objective requires a student to memorize something, multiple-choice tests should not be used, because students will learn just enough to recognize, not to recall. From your own experience, it should be clear that students study less carefully for a multiple-choice or true-false test, than for a completion or short-answer test. Therefore, it is important that the test items and the format should be like the tasks the student will do on the job.

Unfortunately, many authors have treated format decisions as if they were independent of or orthogonal to the requirements of the objective/job, (Popham, 1978, Wesman, 1971, Tinkleman, 1971). It is clear from the task/content classification scheme that format decisions are not arbitrary and must be made with the objective/job in mind.

Classification and Test Development

The classification system can also be used to help determine the number of test items required by an objective for both the Remember and Use levels.

At the Remember level, and at the Use level for Procedures, there should ideally be at least one test item for each objective, or, if one objective covers several pieces of information, there should be test items for all pieces. For example, if the student must remember the part names and functions of several pieces of equipment, this requirement could be incorporated in a single objective. But, then there should be enough test items to test all the information specified. Similarly, if the objective requires the student to perform the steps of a procedure, then the test item(s) must assess performance of all steps of the procedure.

At the Use level for categories, rules, and principles, determining the required number of test items is more complicated. Since these objectives require transfer to new instances, there should be enough test items so that it is possible to decide if the student can apply what has been taught to new instances.

For both categories and rules at the Use level, the procedures for developing enough test items are reasonably well worked out. For categories, Markle and Tieman (1969) Merrill and Tennyson (1977) have shown how to build test items that assess categorization decisions on the basis of presence or absence of the "critical attributes" or "critical characteristics" which define category membership. For rules, Durnin and Scandura (1973), and Brown and Burton (1978), have shown that it is possible to analyze a rule to determine the minimal number of items which can thoroughly test the application or use of the rule. The technique involves tracing all possible "paths" through the algorithm or rule, then constructing an item for each path. The number of items can be further reduced if some paths are "nested" within others.

For principles at the Use level, however, the situation is less clear. There appear to be no prescriptive techniques for analyzing principles for test development purposes as yet. Several areas of investigation, though, hold promise. Richards (1979) is concerned with analysis of principles for teaching purposes, and researchers in the areas of text processing inferencing and "qualitative reasoning" are developing methods for analyzing processes which underly explanation and prediction.

"Real World" Constraints on Testing

The previous section discussed the test item development implications of the classification system. These included prescriptions concerning consistency, format, and numbers of items which ideally should always be followed. Unfortunately, "real world" circumstances often result in insufficient budgets for test development, administration, and scoring, and in insufficient time available for testing versus instruction. Such constraints result in testing situations which are less than optimal. In these situations, testing decisions, again, should not be made arbitrarily. A careful analysis of the objective/job requirements and the constraints should be made, and test items should be designed so that the conditions, actions, and standards approximate as closely as possible the requirements of the job. Moreover, it is incumbent on the test developer to document what is being measured, what is not being measured, and what the effect is on the use or interpretation of test results. With these caveats in mind, let us now consider what kinds of constraints occur, and what strategies are available to deal with them.

Constraints:

Two types of constraints occur for tests of remember-level information:

(a) There is not enough time to test all the information required by the objective(s).

(b) There are not enough resources for scoring, so that easy-to-score tests must be used.

Three types of constraints occur for use-level tests:

(a) There is not enough time to test for all aspects of transfer as required by the objective.

(b) There are not enough resources for scoring, so that easy-to-score tests must be used.

(c) Cost or safety prohibits testing the real task.

Dealing with constraints at the remember level:

(a) Not enough time. The obvious strategy is to sample from the content. This can be done in two ways:

(1). If it is possible to prioritize the pieces of information to be remembered according to some measure of importance, then the most important ones are tested. Priorities can be determined by subject-matter experts or from course goals. The prioritizations have direct implications for the number of test items and the cut off score. Low priority topics should not be tested as extensively and should have lower cut off scores than high priority topics. It is possible to have several priority levels. In this case the number of items and cut off score should reflect the priority assignment. When topics are prioritized, it is then necessary to design the instruction so that students know what will be tested, and how to study for the test. This means that the number of practice items and the instructions to the student about what and how to learn should also reflect the prioritization.

(2). If all the information is important and cannot be prioritized, then pieces of information must be sampled essentially at random. Even here, however, it is usually possible to group the information in related topic areas and sample from all topic areas.

In this case, it is also necessary to design the instruction so that students know how they will be tested and how to study. Here, though, students should be told that all information is important to learn, and that any piece of information may be on the test. Practice should be designed to lead the student to study the type of information to be tested, not just the specific information in the practice.

(b) Easy-to-score tests. This type of constraint often requires that test designers base format decisions on ease of scoring rather than on appropriateness of the item format for the objective being tested. The typical practice is to use selected-response items rather than constructed-response items. It is important to recognize that this format change forces changes in the actions (behavior) being tested, or in the conditions of performance.

There are two strategies which should be followed when dealing with this constraint:

(1). Practice items should be constructed so that they are consistent with the original intent of the objective. That is, practice should be constructed-response even though the final test is selected-response.

(2). Test items should be constructed to measure comprehension or understanding rather than verbatim recognition of information presented in the instruction. When possible, the student should be required to "recall the meaning" of the relevent information. This can be accomplished through the careful use of paraphrasing. An article by Anderson (1972) describes systematic procedures for doing this.

Dealing with constraints at the use level:

(a) Not enough time to test all aspects of transfer. For categories and rules, the methods described earlier (Markle and Tieman, 1969; Merrill and Tennyson, 1977; Durnin and Scandura, 1973) can be used to minimize the number of items required. Therefore, using these methods, it may be possible to test the objective completely even within the time constraints imposed.

In cases where these methods are not applicable, the task should be analyzed and the aspects of transfer most often encountered on the job should be idendified. The test should emphasize those aspects. In addition, if there are any critical (health, safety, etc.) aspects to the task, they should also be tested. Also in these cases, the student should be made aware that there may be transfer situations on the job which he may not be able to deal with.

(b) Easy-to-score tests. Since most item formats can be appropriate for the use level, this situation occurs only infrequently. It occurs when the on-the-job task is a "constructed-response" type of task (for example, in a categorization task when the category names must be produced rather than selected). If constructed-response items cannot be used on the test, they should be used during practice, and, if possible, selected-response test items should use paraphrases according to Anderson's (1972) rules.

(c) Cost or safety prohibits testing the real task. It is often impossible or not economically feasible to test the actual task to be performed on the job. In this situation some type of simulation is necessary. It must be recognized that any simulation situation involves some deviation from the conditions, standards, and/or action of the objective. In general, it is most important to preserve the action from the objective in the simulation.

Simulations are only required in situations where equipment is involved in performance of the task. These tasks generally have three components: perceptual, motor, and cognitive. That is, the student needs to find his way around the equipment, to manipulate the equipment physically, and to know what to do with it.

The first step in deciding on the simulation requirements is an analysis of the task to determine which of the components a most important in task performance. Some tasks, for example, have relatively trivial perceptual and motor components, and the essential part of the task is the cognitive part. (Electron troubleshooting is a good example.) In this case, paper-and-pend simulations are appropriate, and are probably all that is really necessary (McGuire, Solomon, & Bashook, 1976). Other tasks, however, involve complex perceptual-motor responses.

Simulations of these types of tasks should preserve the conditions of the objective that involve perceptual-motor complexity, and should preserve the standards of the objective that are necessary for assessment of the performance.

Summary:

This paper describes a system for classifying objectives that is reliable (Wood, Ellis, and Wulfeck, 1978) and which has specific prescriptive implications for criterion-referenced test item development. Some prescriptions and their underlying rationales were discussed. Finally, "real-world" constraints on the testing process were analyzed in terms of the classification system, and strategies for dealing with them were identified.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research 1972, 42 (2), 145-170

Briggs, L. D. Instructional Design: Principles and applications. Englewood Cliffs, N.J.: Educational Technology Publications, 1977

Brown, J. S., & Burton, R. R. Diagnostic models for procedural bugs in Basic Mathematical Skills. Cognitive Science, 1978, 2, 155-192.

Durnin J. H., & Scandura, J. M. An Algorithmic approach to assessing behavior potential: Comparison with Item forms and hierarchical Technologies. Journal of Educational Psychology, 1973, 64, 262-272.

Gagne', R. M. The content analysis of subject matter: The computer as an aid in the design of criterion-referenced tests. Instructional Science, 1976, 5, 1-28.

Markle, S. M., & Tiemann, P. W. Really understanding Concepts. Champaign, IL.: Stipes, 1969.

McGuire, C. H., Solomon, L. M., & Bashook, P. G. Construction and use of written simulations. New York: Psychological Corp., 1976.

Merrill, M. D., & Tennyson, R. D. Teaching concepts: an instuctional design guide. Englewood Cliffs, N.J.: Educational Technology Publications, 1977

Popham, W. J. Criterion-Referenced Measurement. Englewood Cliffs, N.J: Prentice-Hall, 1978

Richards, R. E. Principle Learning. Unpublished manuscript, 1979.

Tinkleman, S. N. Planning the objective test. in R. L. Thorndike (Ed.) Educational Measurement (2nd Ed.). Washington, D. C.: American Council on Education, 1971.

Wesman, A. G. Writing the test item. in R. L. Thorndike (Ed.) Educational Measurement (2nd Ed.) Washington, D.C.: American Council on Education, 1971

Wood, N. D., Ellis, J. A., & Wulfeck, W. H. Instructional Strategy diagnostic Profile Workshop Evaluation (NPRDC SR 78-17). San Diego: Navy Personnel Research and Development Center, 1978

THE EFFECTS OF TEST-ITEM TYPE
ON
LEARNING AND RETENTION

Persis T. Sturges, Kathleen A. Lockhart
Nicholas H. Van Matre, and Judith Zachai

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

Computer Managed Instruction (CMI) has been implemented throughout much of the Navy's basic training curriculum, resulting in more efficient handling of the large numbers of students who pass through these courses each year. The system allows considerable individualization of instruction, including self pacing and remediation assignments geared to students' problem areas. All testing materials (other than laboratory and performance tests) use multiple-choice as the response medium, and students' answer sheets are machine scored. In view of the number of tests administered daily in the average learning center, this represents a considerable savings in time for the instructors, allowing them to perform other critical instructional functions such as counseling and tutoring students and monitoring their progress in the course.

The hardware currently available in the Navy computer-managed instruction (CMI) system precludes the use of other forms of test materials such as short answer or fill-in (constructed response) items, at least so long as tests are to be machine scored. The administrative personnel of the Propulsion Engineering (PE) School at Great Lakes Naval Training Center, however, have expressed concerns about multiple-choice that generally focus on the issues of learning and retention. Since multiple-choice items require only recognition rather than recall, some school personnel feel that students will not learn the material adequately and therefore will retain less information. To remedy the perceived deficits of multiple-choice testing yet continue to work within the constraints of the computer scoring capabilities of the CMI system, school personnel have developed a conversion procedure that permits use of a constructed response format. Students in the PE course take constructed response module tests, then convert their answers to multiple-choice alternatives presented on a choice list of answers for each question. These new answer sheets are then computer scored. This procedure is time consuming and involves some risk of inaccurate test scores should students make errors during conversion, but it does represent a good faith attempt by school personnel to maximize the effectiveness of their course without sacrificing the benefits of the CMI system. Still, analysis of these conversion tests revealed that while they did require students to write their own answers, 85% of the questions provided cues. While provision of cues for these items is entirely consistent with the objectives as specified for the course, it is not clear to what extent these tests differ from those in a multiple-choice format, and since the procedure has the disadvantages of being cumbersome and time consuming, there is reason to examine the assumption on which it is based. Do constructed response tests offer sufficient advantages over multiple-choice items to warrant the lcss of time and accuracy involved in their use? The research problem, then, was to determine the effects of these two forms of testing on both learning and retention.

METHOD

One hundred twenty students enrolled in the Basics Course of the Propulsion Engineering School served as subjects. Students were assigned non-systematically to one of four learning centers, resulting in four groups of thirty subjects each.

This research investigated three aspects of test item format of the module tests currently in use at the Propulsion Engineering School Basics Course: availability of cues, construction of responses, and conversion of answers. These aspects were systematically varied to form the following different types of test items for use in the module tests and the comprehensive exams.

1. Test items that required the student to write his own response - constructed response (CR) - versus items that required the student to select 1 of 5 alternatives and to indicate the letter corresponding to his choice - multiple-choice (MC).

2. Test items that included cues such as parts lists versus those that contained no cues.

3. Test items that involved the conversion procedure used at the PE School versus those that did not involve conversion.

These test items were used to construct a different type of module test for each of the four experimental groups.

Table 1

Characteristics of Test Formats for Each Group

Group	Cues	Response Format	Conversion
A	Yes	Constructed	Yes
B	Yes	Constructed	No
C	No	Constructed	No
D	Yes	Multiple-choice	No

Group A had the tests and testing procedure currently in use at the Propulsion Engineering School. More than 85% of the items presented cues; the student constructed his response which he converted to the computer answer sheet for computer scoring. The conversion sheet listed five alternative answers for each item number (the 5th alternative was always "None of the Above"), but it did not include item stems. The student matched his constructed response to the alternatives listed and transferred the closest approximation to the computer answer form.

Group B had the tests with cues. The student constructed his response, but the tests were scored and the computer answer form was prepared by the experimenters. The student then submitted this answer sheet for computer scoring. This procedure provided a check on frequency of student conversion error.

Group C had tests with less than 5 percent of the items presenting cues. The student constructed his response, and the experimenter scored the tests and prepared the computer answer sheets. The module tests for Group C were identical to those for Groups A and B except that Parts Lists and other answer choices were eliminated wherever possible without destroying the meaning.

Group D had multiple-choice tests, and students responded directly on the computer answer form for computer scoring. The multiple-choice test was constructed by taking the stem from the tests in use at the Propulsion Engineering School and the five alternatives from the conversion sheet.

Table 2 presents examples of a test item from each of the different test formats. Two series of all tests were constructed so that students requiring repeated testing took the second test from the alternate series with the same format.

Dependent Variables

Several aspects of students' performance served as dependent variables. These were:

1. Learning as measured by:

   a. Students' performance on a comprehensive test composed of each item type.

   b. Students' gain in score from a pre-course administration of a PE knowledge test to a post-course administration of the same instrument.

2. Knowledge retention, as measured by a second administration (two weeks following course completion) of the comprehensive test.

3. Time factors in the course.

   a. Time spent taking module tests

   b. Total time required for conversion

   c. Total time to complete the course.

Table 2

Examples of Item-Type for Each Group

GROUPS A & B: CONSTRUCTED RESPONSE WITH CUES

1-2-2
Complete the following statements.

49. When using a wrench to keep from skinning your knuckles, ____________
the wrench ________________ you. (pull/push)
(toward/away from)

GROUP C: CONSTRUCTED RESPONSE WITHOUT CUES

1-2-2
Complete the following statements.

49. When using a wrench to keep from skinning your knuckles, ____________
the wrench ____________ you.

GROUP D: MULTIPLE-CHOICE

1-2-2
Complete the following statements.

49. When using a wrench, to keep from skinning your knuckles you should
__________ the wrench __________ you.

1. pull, away from.
2. push, away from.
3. pull, toward.
4. push, toward.
5. none of the above.

RESULTS

The results of this study focused on the effects of test item format on learning, retention, and time to complete the course. Supplementary results centered on scoring error frequencies by item format. Each analysis consisted of analysis of variance and, when appropriate, up to three a priori planned orthogonal comparisons:

1. Group A versus Group B to test for effect of conversion. (This comparison contrasted constructed response tests with cues with and without conversion.)

2. Groups A & B versus Group D to test for effects of test format. (This comparison contrasted tests with cues and constructed responses versus tests with cues and multiple-choice responses.)

3. Groups A, B, & D versus Group C to test for the effects of tests with cues versus tests without cues.

Findings

Results were as follows: (1) There were no differences in learning (as measured by a comprehensive test and by a post-test) as a function of testing procedure. (2) Retention (as measured two weeks after the administration of the first comprehensive test), however, was greater for the group (C) that received the "constructed response without cues" test. The simple multiple-choice, the constructed response with cues, and the conversion groups did not differ in retention. (3) The group that received the "constructed response without cues" test took more time to complete the course than did the other groups. This latter group rated their tests as more difficult than did any other group. (4) Scoring error frequencies did not differ by group.

SUMMARY AND DISCUSSION

1. These results indicate that there was no effect of the four test item formats on the basic measures of learning at the end of the 13 module tests. All groups, regardless of test condition, scored equally well on the post-test and the initial comprehensive test.

2. On the measures of retention there was no effect of the different test item formats on the actual score on the second Basic Comprehensive Test, but there was an effect on the amount of loss over the two-week period. The group (C) that took module tests without cues showed less loss on items with cues than the other groups whose module tests had all presented cues and who therefore had practice on this item type. This result suggests that retention can be improved when test items require more than the objectives specify, and suggests that future research is warranted to determine optimal ways in which training and tests can demand more than what is specified in objectives; it would be expected that the amount and kind of additional training demands would differ with the type of task(s) involved in the objectives.

3. The type of test currently used by the PE School (constructed response with cues and conversion) generated no better learning and retention than did the multiple-choice test on any of the criterion test item formats. Further,

this group spent 4½ hours/student in conversion time. Thus, the procedure provides no gains in student performance, but adds costs in terms of time. This conclusion is supported by the fact that the alternatives for the multiple-choice items were derived from the conversion sheets used by the PE School and should not be considered optimal alternatives for multiple-choice items. Using students' most frequent errors to construct alternatives for each question or listing choices that require fine discrimination might be expected to result in even better retention after multiple-choice module tests.

4. The group (C) showing better retention (constructed response without cues) took more time to complete the course. This group did not take more tests (including retakes) but they spent more time taking tests, as well as in other activities during the course. Anecdotal data suggest that instructors and students in this group felt they were being subjected to unusual treatment. This factor may in part explain the increased time in the course, although test difficulty was no doubt a major factor.

5. Examination of the students' attitude data indicated that (1) students taking tests without cues (C) rated their tests as more difficult than did the other three groups, and (2) students using the conversion procedure (A) liked their tests less than did the other three groups.

ANALYSIS AND CLASSIFICATION OF PERFORMANCE TESTING PROBLEMS[1]

William Osborn

Human Resources Research Organization
Fort Knox, Kentucky

INTRODUCTION

The controlled observation of job behavior under realistic but standardized conditions is referred to variously as a hands-on test, a job sample test, a job proficiency test, or simply a performance test. Regardless of the label, the character of tests of this kind is the same: the domain of job behavior is partitioned into tasks, and for each task a criterion of performance is defined and the job-relevant conditions are created; the examinee is then given an instruction to perform, and his performance is observed and evaluated against the established criterion. Such tests, if properly developed and reliably scored, have appeal (Siegel, 1972). Whether used for training evaluation or job certification, they seem to have an intrinsic validity - having an examinee demonstrate a criterion behavior is inherently more appealing than having him answer questions about the behavior. Yet, the development and use of performance tests are not without problems.

It is the nature of many job behaviors that task conditions, equipment, time, and cost considerations inhibit, if not preclude, their enactment in a performance testing mode. Moreover, individual test administration, for other than small scale testing programs, is usually seen as prohibitively expensive and time consuming (Harris, 1962). Even where group testing is possible it is seldom feasible to cover all tasks in the domain and do so under test conditions that faithfully represent the realism of the job environment. Task or part-task sampling together with simulation have been proposed as ways of overcoming some of these problems (Glazer, 1954; Frederikson, 1957; McGuire, 1967; Osborn, 1970; Schriver, 1974). But when and how to use these approaches, or whether they are sufficient in overcoming the barriers to performance test feasibility, are matters that have not been studied. The variety, frequency, and severity of such barriers have yet to be documented.

PURPOSE

The purpose of the work reported here was to identify and classify performance testing problems and to examine them briefly for possible solutions. This was done by analyzing in detail a large sample of job tasks from an Army combat job specialty.

[1] This paper is based on research done under Contract No. DAHC 19-74-C-0059 with the U.S. Army Research Institute for the Behavioral and Social Sciences. Conclusions and opinions expressed are the author's and not necessarily those of the U.S. Army.

APPROACH

A sample of 100 tasks was selected from an existing inventory of approximately 700 tasks spanning all skill levels for the job of Reconnaissance Specialist. Tasks were selected proportionally and randomly from each of 31 content areas. The most heavily represented content areas in the initial sample were: Communication (6 tasks), Leadership (6 tasks), Tracked Vehicles (6 tasks), Machineguns (12 tasks), Tactics (6 tasks), First Aid (5 tasks), and Land Navigation (5 tasks).

As an additional precaution to insure a reasonable variety of job tasks, the initial sample was subjected to another screening. This was accomplished using a simple taxonomic structure defined by three levels of task behavior and three classes of task display. The levels of task behavior were taken from Ammerman and Melching (1966) and are characterized as follows:

- Specific Task - A particular work activity, with a clear beginning and ending point, that is performed under a specific set of task conditions.
- Generalized Skill - A relatively specific activity performed under similar but not identical task conditions.
- Generalized Behavior - A manner of behaving or way of doing things; e.g., application of values and principles.

The types of task display - People, Data, Things - were taken from the Department of Labor's occupational analysis work (1965). Combining these two factors produced nine categories into which tasks were sorted.

As expected, classification of the 100 tasks resulted in a high concentration of Specific Tasks and Generalized Skills, chiefly dealing with Things. Virtually absent were Generalized Behaviors. While the distribution accurately profiled the reconnaissance specialist's job, it was not considered satisfactory for our purposes. Therefore, the 600 remaining tasks were again screened specifically for Generalized Behaviors, as well as additional People and Data oriented tasks; that is, the attempt was to over-sample in all categories but those of Specific Tasks and Generalized Skills pertaining to Things. The final set of 100 tasks was distributed as shown below:

		TASK DISPLAY			
		People	Data	Things	
TASK BEHAVIOR	Specific Task	4	11	31	46
	Generalized Skill	11	15	20	46
	Generalized Behavior	2	3	3	8
		17	29	54	100

Some representation was provided in each category, which was about all that could be hoped for given the nature of the reconnaissance specialist's job. Though less than ideal, the variety of tasks was considered adequate for the purpose of identifying the full range of potential testing problems.

Once the working set of 100 tasks was decided on, a fully relevant job performance test was developed in concept for each. In conceptualizing these tests, every effort was made to capture realistic aspects of the job situation in which the tasks would be performed. For example, in the task "Apply first aid measures for a fracture, sprain, or dislocation," consideration was given to: (a) the types and numbers of casualties to which each trainee would be exposed; (b) the stressful environment in which the trainee would be required to perform; (c) the various methods by which the trainees' performance would be scored; and, (d) the time required to test each trainee. With a concept for the fully relevant test in mind, it was then evaluated from the standpoint of feasibility. Feasibility was judged principally in terms of the test's conformity to the following hypothetical but not unrealistic constraints:

- 50 soldiers to be tested
- one hour to complete testing
- three test administrators or monitors
- no more than one item of major equipment per ten soldiers
- go/no-go and process scores required

In addition to problems created by the above constraints, other special features were noted which, if not dealt with by other than a conventional test mode, would severely jeopardize test validity or reliability.

With the assistance of a military subject matter specialist each of the 100 job tasks was explored in this manner, and the problems recorded by task in a narrative fashion. The full list of problems was then reduced by collapsing similar problems into categories. Finally, the problem categories were organized and redefined in terms of general task characteristics or dimensions.

RESULTS

Two major dimensions of job-tasks, Task Conditions and Task Behavior, were identified as potential sources of difficulty for the test developer in achieving fully relevant yet feasible tests. Outlined as follows is the structure of task characteristics which inhibit the development of performance tests, and which must be dealt with through simulation, task element sampling, or other means if near optimal compromises between validity and feasibility are to be achieved. The type and relative frequency of these testing problems is summarized in the table below.

Task Conditions

Scarce. A large percentage of tasks involved task-relevant conditions (equipment, terrain or support personnel) that are costly or otherwise difficult

FREQUENCY AND TYPE OF TESTING
PROBLEMS ANTICIPATED IN A SAMPLE OF 100 JOB TASKS

Problem		Frequency	Sample Task
Task Conditions			
Scarce		84	
Equipment/Facility	62		• Zero 551 gun launcher
Terrain	43		• Navigate with a compass
Personnel	34		• Lead a security patrol
Dangerous		33	• Prepare a shaped chnrge
Variable		38	
Surround	5		• Engage target under limited visibility
Display	34		• Camouflage a weapon
Latent		12	• Give CBR alarm
Task Behavior			
Long Process		18	• Conduct marksmanship training
Transient Process		43	• Communicate by radio
Affective		12	• Maintain light discipline

to obtain in the quantity normally required to efficiently test a large number of personnel in a reasonable length of time. Of the sample of 100 tasks examined, 62 involved equipment and facilities which would normally not be available in sufficient quantity for any type of group testing. Because of terrain requirements, full-field testing on 43 of the tasks would be difficult or impossible to carry out on other than an individual basis. Similar testing limitations would exist for 34 of the tasks which involve either large numbers of specially trained personnel to participate as task-relevant conditions, or support personnel to act as test controllers. A total of 84 tasks require one or more of these three types of scarce resources.

Dangerous. Thirty-three tasks were identified as having conditions which, if realistically created, would be either physically or psychologically hazardous to people being tested. In nearly all of these cases the conditions are sufficiently dangerous to unequivocally preclude testing with full realism.

Variable. Standardization of conditions is fundamental to valid testing, yet 38 of the tasks present some problem with respect to maintaining control over task conditions. Five tasks involve aspects of the surround (e.g., wind and visibility) which would be most difficult to reproduce from one test session to another. In 34 tasks, elements of the display presented control problems. Thirteen cases involved people as task-relevant aspects of the display; and the behavior of others would be difficult to standardize over test administrations. Other display conditions which would present standardization problems were identified in 21 tasks (e.g., materials available for camouflaging a weapon, or the condition of equipment to be serviced).

Latent. For lack of a more descriptive term, the word "latent" is used to indicate task relevant conditions which require detection as well as immediate reaction by the person being tested. Tasks with this feature may complicate the testing process, in that a prolonged time period may be needed to accommodate the vigilance set required. Only 12 such tasks were identified, including "Give CBR alarm," "Perform duties of an interior guard," and, "Avoid poison plants." Testing problems created by tasks of this type are similar to those encountered for affective task behavior mentioned below.

Task Behavior

Lengthy Process. The time required to execute a task is an obvious factor to be considered in developing feasible tests. Eighteen tasks were judged to take more than an hour. In most of these cases the testing time was estimated at from two to four hours. However, in the case of a task like "Recommend personnel for promotion," full enactment of the task could well involve several days or even weeks if even minimally realistic conditions for personnel data collection were created.

Transient Process. In achieving efficient tests, perhaps the most constraining task characteristic is that of a transient task process. The process in performing some tasks is preserved in or inferrable from the product or outcome of the task (e.g., "Prepare a written message," or "Camouflage a weapon"). Tasks of this sort, unless prohibitively expensive equipment is involved, can be group tested. On the other hand, tasks in which process is not preserved in the product (e.g., "Communicate information over radio net" or "Ground guide a wheeled vehicle") must be administered on an individual basis in order for the tester to record performance. A variation of this problem occurs in some task products which, were it not for certain critical safety precautions in task performance that must be observed, otherwise preserve the process (e.g., splinting a broken leg or disassembling a weapon).

Affective. Twelve tasks were seen as having a substantial affective or "willingness to perform" component. This means that some unobtrusive method of testing would be called for. In testing on the task "Maintains light discipline," for example, the soldier must not know he is being tested if the test is to provide a valid measure of task performance. In the case of unscheduled preventative maintenance tasks on critical equipment (weapons), deciding to perform the maintenance is at least as important as being able to perform it. Realistic enactment of conditions for tasks of this sort presents problems from the standpoint of creatively using testing time to unobtrusively permit the behavior to be emitted.

DISCUSSION

Having identified these problem areas, the relevance of various testing approaches involving simulation or task-element sampling may now be considered.

Dangerous Task Conditions. The problem area of highest priority, perhaps, for use of simulation in testing is that in which the creation of realistic task conditions constitutes a danger to the soldier being tested. In these instances a fully relevant performance test is totally and unequivocally out of the question, and some type of simulation is a necessity. This has long been recognized both as a training and testing problem. On one hand, such tasks are usually the most important to train and test effectively, for the very reason that they are dangerous; on the other hand, attempts to realistically generate a sense of danger, through any means, constitute unethical if not immoral treatment of participants. Severe threat to life, limb or psychological stability must be avoided in training and testing situations. Yet simulation can offer a means of resolving this dilemma.

To be effective, a simulation does not have to duplicate or even approximate the actual stressful stimulus. It must provide a sufficient level of stress to instill a sense of psychological intensity or urgency on the part of the person being trained or tested, and its form should be such that it effectively duplicates the real stimulus in the control exerted over the criterion behavior.

A good example of this is seen in an innovative simulation being used in an infantry platoon combat exercise (Shriver, 1975). The principal feature of the method involves each soldier having a number on his helmet and an inexpensive scope mounted on his rifle; then, during the exercise one soldier may "shoot" another by spotting and correctly reporting the other's number, or "be shot" by allowing his number to be sighted by the other. Number size and scope power have been carefully calibrated from empirical data so that the probability of a simulated kill is highly correlated with the expected outcome in actual battle. The appeal of this simulation is that, without real bullets, the situational feedback exerts realistic control over players' behavior. Experienced soldiers report that mistakes leading to "death" in the simulated situation are virtually the same as those that cause people to be killed in combat.

Similar effects could be created for other dangerous tasks. An effective simulation in "Preparing a shaped charge," for example, might involve a clay substitute for the plastic explosive with a buzzer assembly attached. If the sensitivity of the buzzer were closely calibrated to that of the actual explosive, handling the mock explosive could be realistically tested.

The point here is that the simulation need not entail anything physically similar to the aversive stimulus, or even psychologically similar in intensity of its threat. The substitute stimulus threat must only enable a timely, accurate, unequivocal, and public record of response adequacy. The importance of carefully designed, effective simulations for tasks of this sort is apparent in light of the fact that resulting test conditions constitute the most relevant criterion situation - there neither is nor will be a better criterion against which to validate the simulation. Only in an actual war environment will soldiers face real bullets or deal with live explosives.

Scarce Conditions. Unavailable equipment, facilities or terrain represent the traditional areas for use of simulation. The fidelity of equipment simulations required for effective training has been the object of much study (e.g., Dougherty, 1957; Cox, 1965; Prophet, 1970). A result of possible significance to test development is that relatively low level simulations are adequate for training simple procedural tasks, while high fidelity is needed for highly skilled tasks. This generalization applies equally well to motor and perceptual skills. Though the latter may be less well researched, there is evidence for discrimination or recognition tasks that extremely high fidelity stimuli are necessary for training to transfer positively to the field setting (Mackie, 1964; Baldwin, 1973).

The same notion may be tentatively generalized to problems of unavailable terrain. Where task behavior involves fine-grained perceptual discrimination pertaining to terrain, as in tactical driving or target identification, very high fidelity terrain simulations will likely be required. On the other hand, where task performance is relatively robust with respect to perceptual feedback from terrain features - as in conducting an administrative movement, breeching a minefield, or possibly even in land navigation tasks - it would seem that fairly degraded terrain simulations could be used. There is evidence to suggest that, where miniaturization has been found appropriate in training, the degree of reduction and fidelity of terrain simulations are not highly significant factors (Baker, 1964).

The problem of personnel as relevant conditions in task performance actually breaks down into two separate issues. First, there is the ease with which people function as reasonably reliable, methodical and responsive elements of the environment, much like equipment or other inanimate task-relevant conditions; examples include, "conduct an administrative movement," "coordinate unit defense plans with adjacent units," "supervise a route reconnaissance." In such instances "personnel-objects" can probably be modeled without greatly reducing effective fidelity. The second class of task involving people as task-relevant conditions is that which requires skilled and adaptive interaction, whether verbal or physical, between the person being evaluated and the object person. As this second case centers around the problem of variable task conditions, it is discussed below.

Variable Task Conditions. Difficulties in controlling conditions for certain tasks present problems that certainly should be addressed through simulation. As with dangerous conditions, though to a lesser extent, variable conditions seriously inhibit good performance testing at most any cost. It is usually a two-sided problem. If one attempts to control conditions by testing people individually in exactly the same setting and with identical conditions, wear and tear from repeated administrations begin to leave telltale signs that prompt or mislead subsequent test subjects. On the other hand, it may not be possible to circumvent that problem by creating several identical sets of conditions. Conditions for "camouflaging a weapon" or "clearing fields of fire" simply cannot be duplicated time and time again without giving up some degree of standardization or realism. Similarly, creating a "standard" amount of dirt, rust, wear, and operational deficiency in weapons so that soldiers may be tested on their maintenance-services performance is simply not feasible. If in these cases one can assume that the tasks require minimum motor skill, and it is the perceptual and decision skills that predominate, the effective use of simulation is possible.

The issue of people as variable task conditions was mentioned. Tasks like "apply psychological first aid," "investigate complaints," and "recommend personnel for promotion" require demonstration of interpersonal skills ranging from empathy to actual physical intervention. Here, the other person or group of people represent an important part of the environment to be controlled - that is, standardized - from one test administration to the next. People are difficult to standardize. And other than through the use of well trained stooges or "standardized others" in a role playing mode, effective solutions through simulation are not obvious. This is an interesting and important probler for research.

Long Task Process. Where task performance is extremely time consuming, the concept of part-task testing would seem particularly useful. In navigating from point A to point B on the ground using map and compass, for example, a soldier might be tested on three elements: (1) set an accurate compass heading, (2) pace exactly ten meters, and (3) calculate the number of paces necessary to arrive at point B. This would eliminate the time required to actually negotiate the full distance on foot, and presumably without significant loss in test validity. Breeching a minefield could be similarly tested by requiring the soldier to designate a route through the field and demonstrate his ability to probe for a mine.

Latent Conditions and Affective Behavior. Testing problems associated with tasks requiring an unalerted reaction to infrequent and unpredictable stimuli do not seem amenable to solution through simulation or task sampling approaches. The same holds for tasks with a substantial affective ("will do" as opposed to "can do") component. In either case, effective test methods must center around creation of an unalerted and unobtrusive set; that is, the soldier must not be aware that he is being tested, or at least unaware that he is being tested on that particular task. Solutions to this problem lie in an approach that, (a) satellites the task on another task that is ostensibly the one being tested, or (b) embeds the subject task in a job context of other tasks that are being tested as a functional module (Osborn, 1974).

Transient Task Behavior. Tasks consisting of elements or subtasks that are not scorable by observing just the task product or outcome present a problem that also is not solvable directly through simulation or sampling approaches. To avoid the requirement for a tester to observe and score each soldier's task performance as it is executed, performance may be recorded by audio means (e.g., "communicate information over radio net") or by video means ("ground guide a wheeled vehicle") for scoring later. Should competent test administrators not be available, or other problems associated with individual testing prevail, the use of recordings may be easily justified. Where diagnostic scoring is not required, this problem of course can be ignored.

CONCLUSION

This paper has presented an analysis of job-task characteristics that are problematic to efficient performance testing. The implications of simulation, task-element sampling and other variations in test procedure were discussed as solutions to the problems. Whether an acceptable balance of realism, standardization and cost can ever be achieved in all areas of performance testing is questionable. Valid tests for some job tasks simply may not be worth the price. For others, however, innovative techniques may be found which enable valid testing at managable cost.

REFERENCES

Ammerman, H.L. & Melching, W.H. The Derivation, Analysis and Classification of Instructional Objectives, HumRRO Technical Report 66-4, May 1966.

Baker, R.A., et al. Development and Evaluation of Systems for the Conduct of Tactical Training at the Tank Platoon Level, HumRRO Technical Report 88, April 1964.

Baldwir, R.D. Capabilities of Ground Observers to Recognize and Estimate Distance of Low-Flying Aircraft, HumRRO Technical Report 73-8, March 1973.

Daugherty, D.J., et al. Transfer of Training in Flight Procedures from Selected Ground Training Devices to the Aircraft, Technical Report NAVTRADEVCEN 71-16, September 1957.

Frederiksen, N., et al. "The In-Basket Test," Psychological Monographs: General and Applied, 1957, 71, (9, Whole No. 438).

Glazer, R., et al. "The Tab Item," Educ. & Psych. Meas., 1954, 14, 283-293.

Harris, A. & Mackie, R.R. Factors Influencing the Use of Practical Performance Tests in the Navy, ONR Technical Report 703-1, August 1962.

Mackie, R.R. & Harabedian, A.A. A Study of Simulation Requirements for Sonar Operator Training, Technical Report NAVTRADEVCEN 1320-1, March 1964.

McGuire, C.H. & Babbott, D. "Simulation Technique in the Measurement of Problem Solving Skills," J.Educ. Meas., 1967, 4, 1-10.

Osborn, W.C. "An Approach to the Development of Synthetic Performance Tests," HumRRO Professional Paper 30-70, December 1970.

Osborn, W.C., et al. "Functionally Integrated Performance Testing," paper for the 16th Annual MTA Conference, Oklahoma City, Oklahoma, October 1974.

Prophet, W.W. & Boyd, H.A. Device-Task Fidelity and Transfer of Training: Aircraft Cockpit Procedures Training, HumRRO Technical Report 70-10, July 1970.

Shriver, E.L. & Foley, J.P., Jr. Evaluating Maintenance Performance: The Development of Graphic Symbolic Substitutes for Criterion Referenced Job Task Performance Tests for Electronic Maintenance, AFHRL-TR-74-57 (III), 1974.

Shriver, E.L., et al. REALTRAIN: A New Method for Tactical Training of Small Units, U.S. Army Research Institute Technical Report S-4, December 1975.

Siegel, A.L., et al. Some Techniques for the Evaluation of Technical Training Courses and Students, AFHRL-TR-72-15, 1972.

U.S. Department of Labor. Dictionary of Occupational Titles, 3rd edition, Vol. II, Government Printing Office, Wasnington, D.C., 1965.

SOME FACTORS THAT AFFECT RELIABILITY OF HANDS-ON TESTS

Patrick Ford and Charlotte H. Campbell

Human Resources Research Organization (HumRRO)
Radcliff, Kentucky 40160

INTRODUCTION

Hands-on testing in the Army gets a good press. All MOS producing schools feature hands-on testing at the end of the cycle. The Skill Qualification Test (SQT) program emphasizes hands-on tests. Some commanders occasionally complain about the cost of administering the tests, but even they are usually more willing to base decisions on results of hands-on tests than written tests. The high acceptability of hands-on tests is due largely to the similarity between job conditions and test conditions.

Although many users of hands-on tests are willing to accept hands-on results uncritically, developers of hands-on tests generally concede that apparent similarity is not sufficient for assuring meaningful decisions based on the test. The SQT system, for example, calls for tryouts with experts and representative soldiers scored simultaneously by two to four scorers and careful review by test specialists. Our office has been involved with developing review procedures for test specialists. During this involvement, two questions have been persistent concerns:

- Is the demonstration of interrater agreement sufficient for assuring classification reliability?

- What item defects should a test reviewer look for to assure classification reliability?

Answering these questions requires data from a reliability study that is more directly related to classification reliability than an interrater agreement study. Classification reliability in this context means the extent to which the grouping of soldiers as performers or nonperformers on a task will be the same regardless of who scores the tests or when the soldiers are tested. Interrater agreement is certainly essential, but it tells nothing about the control of essential test conditions or the stability of the behavior in the soldiers classified. Since test-retest agreement is affected by differences in essential conditions and changes within the soldiers between test administrations as well as differences between scorers, it is a more direct measure of classification reliability than interrater agreement is.

METHOD

The data for this study were results of test-retest administrations of hands-on tests of Armor Tasks. The data were collected as part of an ARI study of retention of ability on Armor crewman tasks. The design for that study called for administering the pretest twice to each soldier. The hands-on tests which covered 52 tasks had been constructed by a contractor. The tests had brief set-up instructions and no specific scoring instructions. They had been reviewed for subject matter accuracy, but had not been revised based on tryouts. Interrater reliability had not been demonstrated.

The subjects for the retention study were 66 recent graduates of OSUT in Armor. The scorers were three Armor NCOs.

Each soldier was tested on a group of tasks, then retested by a different scorer. Retesting, though not immediate, occurred on the same day as the first test.

Fourteen tasks were selected for analysis in this study. Because of the interest in item characteristics, one staff member screened the test results to identify tasks with the most difference between first test and retest in the middle range of raw scores. This screening eliminated tasks on which most soldiers passed or failed all items both times.

Five tasks were selected as a basis for deriving characteristics of items with low test-retest agreement. An agreement index was calculated for each item as the number of scoring agreements between test and retest (the number of examinees who passed both times or failed both times) expressed as a percentage of the total number of examinees. Staff members then examined each item with an agreement index of 69 or lower to determine possible sources of the low agreement. All but one of these items had one or more of the four defects listed in Table 1.

To verify the practicality of relying on reviewers to identify defective items, two staff members then classified each item in the remaining nine tasks to see whether the level of test-retest agreement could be predicted by examining the items in terms of the identified defects. Both staff members are experienced in developing and reviewing hands-on tests. One is a subject matter expert in Armor. Each reviewer independently predicted test-retest agreement for each item as high, marginal high, marginal low, or low. The reviewers then resolved disagreements. There were 16 disagreements on the 87 items for an agreement level of 82%. Neither reviewer had seen the pass-fail data or test-retest data for any of the tasks. The reviewers also decided the nature of the defect in terms of the categories derived from the initial review of five tasks.

During the verification phase several doctrinal errors were noted. For example, "Pulled barrel extension to the rear" is listed as an item in the task, "Apply immediate action to M219 machinegun," but is entirely irrelevant to performing the task. Since the reviewers were unable to agree on the likely effect of incorrect doctrine on test-retest agreement, it was decided to analyze the data under three conditions for the incorrect doctrine items--rated high, rated low, and discarded.

TABLE 1

DEFECTS THAT CHARACTERIZE LOW TEST-RETEST AGREEMENT ITEMS

Defect	Definition	Example
Unobservable action	Scorer cannot see action and is not advised to evaluate any product of action.	Observed for lighting of circuit tester bulb each time Gunner or TC announced ON THE WAY.
Undefined judgment	Item allows various interpretations.	Squeezed magnetic brake switch and rotated Gunner's control handle to traverse turret.
Uncontrolled condition	No assurance that essential test conditions are the same for all administrations.	Plugged electrical lead into solenoid.
Unrealistic cue	Step covered by item is not naturally eleicited from test conditions.	Prepared to load another round in case cease fire is not given.

RESULTS

Once the items were classified by the reviewers into the high group or one of the five defect groups, a one-way analysis of variance was performed, using the obtained agreement index as the dependent variable (see Table 2). Those items for which marginal agreements were predicted were not used in the analyses. The comparisons planned in advance were the high group versus the five low groups, and the high group versus each of the five low groups. Because the tests are not orthogonal, the first comparison (high versus all low) was conducted using a multiple t-test; the 5 pairwise comparisons were conducted according to Dunnett's method for comparisons of experimental groups with a control. The comparison involving the incorrect doctrine category involved a two-tail test; the other four comparisons were one-tail, with the high group predicted as having the higher mean.

The analysis of variance yielded an overall F of 2.48, significant at the .05 level (see Table 3). The comparison of the high group to the five other groups was, at best, marginally significant, with a significance of .10. One of the pairwise comparisons between the high group and each of the five other groups was significant, at the .05 level, that involving the items in the unrealistic cues category. The comparison with the unobservable category items approached the .05 level of significance.

An inspection of the group means revealed that the incorrect doctrine category had the highest mean of any of the groups, leading to two additional comparisons for which the Bonferroni t-statistic was computed. The comparison between the high group and the four low groups (not including incorrect doctrine) was not significant; the comparison of the high and incorrect doctrine groups with the four low groups was significant at the .05 level. Despite the dangers involved in performing multiple nonorthogonal tests, it is clear that incorrect doctrine, as detected by the reviewers, does not lead to low scoring agreement.

A second analysis of variance was performed, with the incorrect doctrine items reclassified with the high group. The overall F is significant at the .05 level (Table 4). Again, the comparisons between the high group and the four low groups, and the high group with each of the four low groups, were planned. And again, the high group items were significantly higher than the low groups items ($p < .05$). The high group was also significantly higher than the unrealistic cues group ($p < .01$) and the unobservable items group ($p < .05$).

Finally, to avoid taking advantage of the high incorrect doctrine agreement indexes, a third analysis of variance was conducted, with the same planned comparisons as in the second. The overall F was again significant (see Table 5) and the high group was significantly higher than the combined low groups and the unrealistic cues group.

TABLE 2

MEANS AND STANDARD DEVIATIONS OF AGREEMENT INDEXES FOR PREDICTED HIGH AND LOW (DEFECT) GROUPS

	GROUPS			
	High	High, including Incorrect Doctrine	Low (Defect)	Low, including Incorrect Doctrine
Number of Items	40	45	32	37
Mean Agreement	73.0	73.6	63.1	65.2
Standard Deviation	14.5	14.1	17.3	17.2

	LOW (DEFECT) GROUPS				
	Unrealistic Cues	Unobservable	Undefined Judgment	Uncontrolled Conditions	Incorrect Doctrine
Number of Items	13	9	5	5	5
Mean Agreement	60.5	60.3	74.0	63.7	78.8
Standard Deviation	21.8	16.4	12.2	4.6	9.1

TABLE 3

ANALYSIS OF VARIANCE AND COMPARISONS FOR SIX GROUPS (HIGH AND FIVE DEFECT GROUPS)

Source	SS	df	MS	F	P
Groups	2995.51	5	599.10	2.484	$< .05$
Error	17127.38	71	241.23		
Total	20122.89	76			

Comparisons

High vs. Five Defect Groups	$t = 1.494$, $p < .10$	(Multiple t)
High vs. Unrealistic Cues	$tD = 2.52$, $p < .05$	(Dunnett)
High vs. Unobservable	$tD = 2.21$, NS	(Dunnett)
High and Incorrect Doctrine vs. Unrealistic Cues, Unobservable, Undefined Judgment, and Uncontrolled Conditions	$d_{.05} = 10.89$, $d_{Obs} = 11.24$, $p < .05$	(Bonferroni)

TABLE 4

ANALYSIS OF VARIANCE AND COMPARISONS FOR FIVE GROUPS (HIGH INCLUDING INCORRECT DOCTRINE AND FOUR DEFECT GROUPS)

Source	SS	df	MS	F	P
Groups	2846.38	4	711.60	2.966	< .05
Error	17276.50	72	239.951		
Total	20122.89	76			

Comparisons	
High including Incorrect Doctrine vs. Four Defect Groups	t = 2.388, p < .05 (Multiple t)
High including Incorrect Doctrine vs. Unrealistic Cues	tD = 2.696, p < .01 (Dunnett)
High including Incorrect Doctrine vs. Unobservable	tD = 2.351, p < .05 (Dunnett)

TABLE 5

ANALYSIS OF VARIANCE AND COMPARISONS FOR FIVE GROUPS (HIGH AND FOUR DEFECT GROUPS)

Source	SS	df	MS	F	P
Groups	2508.87	4	627.22	2.503	< .05
Error	16793.14	67	250.63		
Total	19302.01				

Comparisons	
High vs. [illegible] Defect Groups	t = 2.121, p < .05 (Multiple t)
High vs. Unrealistic Cues	tD = 2.475, p < .05 (Dunnett)
High vs. Unobservable	tD = 2.167, NS (Dunnett)

In order to determine whether the defect categories were sufficient for describing items with low test-retest agreement, the reviewers later classified all low agreement items, whether predicted low or not, intc the four defect categories. The agreement level for classification by category was 80%. The results of the classification are shown in Table 6. For the 14 tests in this study, a reliability demonstration that was not sensitive to unrealistic cues and uncontrolled conditions would fail to detect 66% of the low agreement items.

Even in retrospect, the viewers could not find defects in the seven low reliability items in Table 7. This apparent lack of defects may have three implications. First, the soldiers may not have practiced the task often enough for the behavior to be stable. This is probably the case with "Set fuze for APERS round." The implication is that tests of tasks with such behaviors should require more than one demonstration before the soldiers are classified. A second possible explanation is that the task goes so quickly that scorers are able only to get a general impression of soldiers' levels of competence and the score recorded for each item means nothing. The overall significance of the analysis of variance does not support this explanation, but the interaction between the speed of the task and a relatively low level of training rather than test quality may account for the low agreement for this segment of "Perform prepare-to-fire checks." The third possibility is that there are other factors not identified that affect reliability of hands-on items.

CONCLUSIONS

The first question that motivated this study was, "Is the demonstration of interrater agreement sufficient for assuring classification reliability?" This study suggests that by itself, interrater agreement is not sufficient. Of the 45 items with test-retest agreement lower than 70, most of the 17 items with unrealistic cues and the 15 items with uncontrolled conditions would not be detected by an interrater agreement study. Nevertheless, if only one reliability study can be conducted, an interrater agreement study is more valuable than a test retest study for refining hands-on tests. The advantage of the interrater agreement study is that the results are easier to interpret because of the narrower focus. Although an interrater agreement study cannot measure performance stability or control of conditions, it also does not assume these factors are present. A test-retest study on the other hand must either assume scorer agreement in order to say anything with certainty about the stability of behavior or the control of contions, or assume stable behavior and controlled conditions to check scorer agreement.

An appropriate supplement for an interrater agreement study is suggested by the answer to the second question that motivated this study, "What item defects should a reviewer look for to assure classification reliability?" The significant F test for prediction of test-retest agreement, although not dramatic, indicates that reviewers can locate defective items if they look for unobservable actions, undefined judgments,

TABLE 6

MEANS AND STANDARD DEVIATIONS OF AGREEMENT INDEXES FOR FOUR LOW (DEFECT) GROUPS (ITEMS CLASSIFIED AFTER DATA INSPECTION)

	GROUPS			
	Unrealistic Cues	Unobservable	Undefined Judgment	Uncontrolled Conditions
Number of Items	17	11	9	15
Mean Agreement	52.8	55.9	50.2	60.4
Standard Deviation	14.6	8.5	12.1	7.3

TABLE 7

ITEMS WITH LOW AGREEMENT, NOT CLASSIFIED INTO FOUR LOW (DEFECT) GROUPS, AND TEST - RETEST RESULTS

	Test-Retest Results (Number of Examinees)			
	Pass-Pass	Pass-Fail	Fail-Pass	Fail-Fail
TASK: Perform Prepare-to-fire checks from Gunner's station.				
On command CHECK FIRING SWITCHES:				
. Turned main gun switch ON.	42	10	11	3
. Rotated main gun manual firing device T-handle.	36	14	8	8
Note: Announced ON THE WAY each time a trigger is checked for the main gun or the manual firing device is actuated.				
. Turned main gun switch OFF.	28	18	13	7
. Turned coaxial machinegun switch ON.	14	15	12	25
. Depressed firing trigger on manual elevating control handle.	15	14	10	27
. Turned coaxial machinegun switch OFF.	14	15	10	27
TASK: Load main gun in response to firing commands.				
. Set range on APERS ammunition fuze when "BEEHIVE TIME" is announced.	16	15	16	19

uncontrolled conditions, and unrealistic cues. The significant comparison of the high group versus the low group with unrealistic cues suggests that the highest payoff activity for a reviewer is to look for items that are not naturally elicited by the test conditions. One approach to maximizing reviewers' effectiveness is to have them supplement rather than duplicate the empirical study. This study is encouraging with regard to the probable benefits of focussing on the defects which the empirical demonstration is least likely to detect.

Finally, it should be noted that, by many criteria, these 14 tests are good. There are fewer unobservable actions and undefined judgments than one would normally expect in hands-on tests that have not been tried out. And even the errors are mainly, in Branch Rickey's phrase, errors of enthusiasm. Most of the unrealistic cue items appear to have resulted from a desire to stretch the limits of the hands-on mode to include crew interaction tasks in the absence of a crew or score combat behavior in a safe environment. Such stretching increases the acceptability of hands-on tests. Most of the uncontrolled conditions items could have been corrected through specific set up instructions. But such instructions, since they increase the bulk of the test and imply that scorers do not know best how to conduct the test, reduce acceptability. Thus classification reliability and acceptability appear to clash. Reduced classification reliability is a hidden cost. But if the reduction leads to poor decisions, the cost is very high.

RELIABILITY: PURPOSE AND USE*

Dr. John B. Meredith, Jr.

Data-Design Laboratories

ABSTRACT

In military training programs, it is desirable to assess trainee performance with respect to technical proficiency in order to plan future training paths for the technicians. To accurately determine trainee proficiency, test instruments need to be both valid and reliable. The validity of test instruments can be described in several regards, (i.e., content validity, criterion validity, construct validity). The interpretation of validity, therefore, is dependent on the type of validity under investigation.

Although the reliability of test results can be estimated by several methods (e.g., split half, test-retest, internal consistency), the norm-referenced interpretation of the resultant reliability coefficients is the same for each.

The purpose of this presentation is to elaborate on the concept of reliability, with respect to "True Score Theory" (Lord and Novick, 1968 and Nunnally, 1975), and to the uses and abuses of trainee and technician standardized test results. Further, the presentation includes a method to improve the reliability of future test scores by using the readily available test item statistics of difficulty and discrimination.

*This paper was presented but, due to its unavailability at the time of printing, only the abstract is reproduced here.

SOME STATISTICAL CONSIDERATIONS IN EVALUATING TRAINING PROGRAMS

Frederick H. Steinheiser, Jr.

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, VA 22333

INTRODUCTION

A thorough evaluation of a training program can help to answer such questions as: Is the problem amenable to a training solution? What training method is most appropriate for the content area and target population? Did students learn? Did the learning result in improved job performance? How might the training program be improved? Are there certain "types" of students for whom the training is more (or less) effective?

These questions lead to two distinct, but related methods for evaluating training effectiveness. First, to what extent does "in class" learning occur? Second, to what extent does this learning translate into on-the-job behavior change? These are the topics of internal and external validity, respectively. High internal validity is achieved when the learning can be attributed directly to participation in the instructional program. Any factors offering alternative explanations for the training results will weaken internal validity. External validity will be high when students' job performance is assessed to be higher than it was prior to training. Internal validity is a necessary but not sufficient condition for external validity. No conclusions about the effects of training on on-the-job performance can be made if there is ambiguity about extraneous factors which might obscure the true instructional effects. We must therefore ask: To what extent is the evaluation design able to detect and control for alternative explanations of the results which might otherwise contaminate the genuine measurement of the instructional program?

The objective of this paper is to emphasize some statistical and experimental design issues which are implicated (and often ignored) in evaluating a training program. The first issue is that of experimental design adequacy, to insure high internal validity. "Extended control group" designs will be offered as a realistic solution, and an example is given. A second issue is concerned with statistical hypothesis testing and the likelihood of making erroneous causal inferences about internal validity, based only upon levels of statistical significance (e.g., via F-ratio) attained. The third and final topic addresses the choice of summary statistics, and for purposes of illustration focuses upon the F-ratio and omega-squared ("magnitude of effect") indices following an analysis of variance. These topics are directly aimed at achieving high internal validity; discussion of external validity is beyond the scope of the present paper.

Extended Control Group Designs

Ideally, a true experimental design should be used to evaluate a training program. At minimum, an instructional (experimental) group, and at least one control group should be used. To test the hypothesis that any observed change has resulted mainly from the instructional experience, we also need to know about the task relevant ability levels of the students in all groups before instruction/training is given.

Traihing is often designed to teach new skills/knowledge to an already experience student/soldier. The new material may be an extension or elaboration of the tasks which he already performs. And, through processes of self-selection and counselor guidance, many will seek positions for which they possess some aptitude, interest, or prior knowledge. Trainees may, therefore, enter a program of instruction with a significant background relative to the course content. While such a background should help the student pass the course, the validity of the evaluation of the training program can be compromised. Clearly, prior exposure to aspects of the course content will influence both pretest and posttest scores.

The level of pretraining ability can be made comparable across groups by (1) using a pre-test to measure the relative skill and ability levels of the experimental and control groups, or by (2) randomly assigning students to the experimental and control groups. Pretesting gives control by providing a direct measure of whether or not prospective students differ on the pertinent variables before the experimental group undergoes the training. The randomization approach foregoes direct measurement in favor of assuming that that any large differences will be evenly distributed across the groups.

A strong argument can be made for the advantages of pretest designs in establishing the initial comparability of students across groups by considering the following factors: (1) The same precision of measurement can be achieved without having to increase the sample size; (2) if randomization is not achieved, it will likely be apparent in the pretest scores; (3) subgroups (e.g., due to sex) may have different pretest scores which need to be considered in integrating the post-training scores; (4) the level and range of student abilities prior to developing training programs can offer guidance in the complexity and duration of the program.

The traditional "experimental vs control group" design is not adequate to achieve the benefits that can be accrued from a true extended control group design which provides measurement of pre and post training scores, and measurement of non-training influences as mentioned above. One common type of extended control group design includes two additional un-pretested control groups. This is called the Solomon 4-group (Solomon, 1949). This design allows for the measurement of the effects of pretests on the student's subsequent performance in the training program and posttest. Indeed, the evaluation process itself (e.g., pretesting) can sometimes jeopardize the evaluation of the instruction program through pretest effects. This effect can occur because the very act of measuring a variable before manipulating it can cancel the possibility of observing the effects of that variable as it would have occurred outside of the experimental setting. When a pretest either provides a learning experience or cues the student to recall past learning, exposure to such a measure may artificially inflate or depress subsequent post test scores. An example of how the Solomon 4-group (a type of extended control group design) was used in a training program evaluation will now be given.

The purpose of the study (Bunker and Cohen, 1977) was to evaluate the success of a basic electricity training program for telephone installers and repairmen. The evaluation process included testing for the possible contaminating effects of pretesting through the following four groups of trainees: (1) pretest, train, posttest; (2) pretest, no training, posttest; (3) no pretest, train, posttest; (4) no pretest, no training, posttest. This Solomon 4-group allows the measurement of the direct effect of pretesting on posttest scores, and the interactive effects of pretesting and training on posttest scores. An additional provision was made in this study to assess any complex interactions between aptitude, pretesting and training which might affect simple pretest main effects and pretest x training interactions...even when the Solomon 4-group design is used.

The subjects were 131 males with comparable company experience. Each was randomly assigned to one of the four conditions. The pretest and posttest consisted of equivalent forms of an objective test of basic electricity knowledge. Scores on a Personnel Test for Industry (Numerical Form A) were used to categorize people within each of the 4 groups into three levels of numerical aptitude.

In the first data analysis, the six sets of means from the Solomon design were compared, in order to evaluate the effectiveness of the training, and to assess the extent of evaluation contamination (lack of internal validity) due to pretesting. Secondly, a 2 x 2 ANOVA was performed on the 4 sets of posttest scores. Third, the posttest scores were partitioned to find out if aptitude factors moderate or mask simple pretest effects or interactions.

Results from the mean comparison analyses and the 2 way ANOVA showed a strong main effect for pretesting, with no evidence of interactions. But the 3 way ANOVA did uncover a significant numerical aptitude x pretest x training interaction. Specifically, the posttest scores of low aptitude people were reduced by exposure to the pretest in the training condition, but were unaffected by pretest exposure in the control condition when no training was given.

Further results showed that medium aptitude subjects were hindered by pretest exposure in the training condition, but were assisted when pretested in the control condition. High aptitude subjects were unaffected by pretest exposure in both of these conditions. Interpretation of these results can most parsimoniously be made in terms of "pretest contamination." That is, higher order interactions with aptitude (or other individual differences) variables can obscure simple pretest effects, even when an extended control group design is used.

Why did pretest exposure have such a differential effect? Medium and low aptitude subjects may have shown the least benefit from training because (1) fear of failure was elicited from difficulties with the pretest, (2) inability to integrate training material with the pretest examples, or both (1) and (2). Interestingly, when medium aptitude students took the pretest but did not have training (group 2), the posttest scores showed an increase over the pretest. But the medium aptitude group which was not pretested (group 4) showed no such facilitation.

These observations, along with the pretest's restrictive effects on medium and low aptitude students, suggest that an overall "training effect" may be rather tenuous. Without controlling for the effects of pretesting and aptitude, the self-biasing aspect of evaluation would not have been detected.

Why should researchers doing military training evaluation consider extended control group designs? Suppose that a simpler experimental-and-control group design has been used in the above study. Quite likely, an erroneous conclusion about training effectiveness would have been reached. For example, the scores of the trainee group may have been suppressed, while those of the control group would have been inflated. The result would be a potentially valuable training program that is mistakenly perceived to be ineffective. It therefore seems reasonable to recommend that pretest influence and personal factors (such as aptitude) be controlled when the effectiveness of new training programs is to be evaluated. Furthermore, different training and evaluation procedures may be of use for students' varying aptitudes. Possibly a pretest for medium and low aptitude trainees could be given as part of the MOS selection procedure, rather than immediately preceding specialty training. A more complete list of extended control group designs may be found in Campbell and Stanley (1963) or Mahoney (1978).

Some Precautionary Notes On Statistical Significance (Hypothesis) Testing

A strong experimental design is a prerequisite to statistical analysis, but the former does not guarantee the adequacy of the latter. In this section, a critique of classical statistical hypothesis testing will be presented. It will be argued that "statistically significant results" following a well-designed experiment can still lead to erroneous inferences about the internal validity of that training program. The last section of this paper proposes a possible solution to this dilemma, using the "magnitude of experimental effect" as a summary statistic which can be computed from ANOVA just like an F-ratio.

Statistical significance is synonomous with statistical rareness. Results are significant statistically if they are expected to occur very infrequently in random sampling under the assumption of the null hypothesis--that there is no difference between the treatment and control group--in the simplest case. The typical question asked in tests of statistical significance is: What is the probability of obtaining a large average difference (call it D') between two samples, if the samples were obtained from the same population? (H_0: the samples were indeed obtained from the same population, so that there would be no difference between treatment and control groups.) Symbolically, the question is: What is $p(D'|H_0)$?

For example, assume that $p(D'|H_0) = .01$. If the question is reversed, we would then be asking what is the probability that the two obtained groups (one which received training, the other did not) were sampled from the same population? But this is the question to which an answer is sought via the p value in conventional tests of statistical significance. Symbolically, this question seeks the value of $p(H_0|D')$. But the value of $p(D'|H_0)$...which we assumed to be =.01...is often used as an answer to

$p(H_o|D')$. This kind of a bidirectional interpretation cannot be made, because classical significance tests yield $p(D'|H_o)$, not $p(H_o|D')$. As Cronbach and Snow (1977, p. 52) state this case:

> A p value reached by classical methods is not a summary of the data. Nor does the p value attached to a result tell how strong or dependable the particular result is...Writers and readers are all too likely to read .05 as $p(H|E)$, "the probability that the hypothesis is true, given the evidence." As textbooks on statistics reiterate almost in vain, p is $p(E|H)$, the probability that this Evidence would arise if the (null) hypothesis is true. Only Bayesian statistics yield statements about $p(H|E)$.

If the importance of the non-reversability of conditional probabilities seems to be just so much word play, consider the following rather morbid example (adapted from Carver, 1978, p. 384). The conditional probability that a person is dead (D), given that he was hanged (H), or $p(D|H)$, will be at least .999. Now let's reverse the states of nature, and inquire about the probability that a person has been hanged (H), given that he is dead (D). In contemporary society, this value will certainly be very small, certainly less than .001. Clearly, it would be outlandish to claim that .999 is the probability that hanging was the cause of death for $p(H|D)$. Yet typically in research, $p(H|D)$ is interpreted as if it was an estimate for $p(D|H)$.

If I find that the difference between the average test score for a treatment and a control group is significant at the .05 level, does this mean that if one were to replicate the experiment 100 times, the observed difference would result in 95 of the replications? No! The probability of replication of $p(R/D')$ is not .95. What we do have is $p(D'/H)$, which is the probability of these data (difference between means) if the null hypothesis is true.

Suppose that two groups of students received different kinds of instruction, and that the mean difference between their final exam scores was statistically significant, $p < .01$ (via t-test of F-ratio, or any non-parametric test). Does this result mean that if the results had been significant at $p < .05$, one teaching method would not be as much better than the other one, as when the mean difference was significant at $p < .01$? Not necessarily. As Carver (1978, p. 386) summarizes, it is erroneous ". . . to interpret the size of the p value as reflecting the degree of validity of the research hypothesis, that is the lower the p value such as $p < .001$, the more highly significant or valid the research hypothesis." Indeed, the ability to proclaim a valid research hypothesis is highly dependent upon multiple replications of well designed experiments (such as extended control group designs).

Statistical vs Practical Significance

The more subjects that are used in an experiment, the more likely are the results to be statistically significant. Hays (1963) offered this warning: "Virtually any study can be made to show significant results if one uses enough subjects regardless of how nonsensical the content might be" (p. 326). Although much has been said about the damages of experimenter bias in the conduct of an experiment (Rosenthal, 1969) and in the design of an experiment (Mahoney, 1978), little attention has been paid to the fact that the experimenter's control over the number of subjects that will participate is also a biasing factor.

The odds that a statistically significant result will emerge are enhanced as sample size increases. This can be easily shown for such commonly used statistics as t-tests and analysis of variance. For example, if we assume the truth of a research hypothesis, that $X_1=60$ with $\sigma_1=4.0$ and $X_2=65$ with $\sigma_2=15$, then perusal of a t-distribution table will show that a sample size of at least 8 is needed for $p<.025$, and 9 for $p<.01$.

The difference between statistical and scientific significance is more than a matter of semantics. A large experiment, in which hundreds of subjects are in each group, might easily yield a small difference between group means, yet be statistically significant at $p<.05$, .01, or .001. A detailed example will be offered later in this section of the paper.

For the researcher/evaluator who chooses a textbook ANOVA model in which to cast his data, it is important to note that the initial sampling assumptions can affect the statistical significance of the results. If a fixed effects model is selected, it is assumed that the levels of the independent variables have been exhaustively sampled. (Test items, or types of test items, for example.) This means that no generalization beyond those levels sampled is intended, or theoretically allowable. The random effects model assumes that the treatment variables have been randomly selected from a very large population of such variables... test items, for example. It is possible to generalize the results from this random sample to the entire population (of items, not people). The mixed effects model allows for both fixed and random variables to be studied in the same experiment, with the results for each factor to be interpreted according to that factor's sampling plan.

The choice of a model does have an impact upon the probability of obtaining the observations under the null hypothesis for each variable (condition, factor). Consider the following evaluation which was conducted for the U.S. Army Military Police School at Fort McClellan, Alabama. Each of 237 students shot a total of 240 handgun rounds from eight different position-distance combinations. There were three repetitions of 80 shots each, at stationary silhouette targets. Within each repetition, five shots were taken, the weapon was reloaded, and five more shots were fired in the adjacent test lane. (Each subject had previously passed a training course with a score of at least 35 hits out of 50 shots.) In the test, 160 trials (2repetitions) were taken on Thursdays, the third was taken on Fridays. The completely crossed design was therefore: A x B x C x D, or 237 x 2 x 8 x 3, or subjects x lanes x tables x repetitions.

Table 1 highlights the results of the ANOVA from this experiment. The first column of F-ratios assumes a mixed model, with B, C, D as fixed factors. The second column of F-ratios assumes that only Tables was a fixed factor. The third F-ratio column assumes that all four factors were randomly sampled from their respective populations. The result is obvious: different ANOVA models produce different F-ratios for null hypothesis rejection, given the same set of data.

TABLE 1. Changes in F-Ratios as a Function of ANOVA Model

Source	d.f.[1]	M.S.	F[2]	F[3]	F[4]
A (Subjects)	236	12.80		3.93****	2.54****
B (Lanes)	1	7.70	7.33****	5.96**	2.26
C (Tables)	7	732.71	385.64****	79.11****	79.11****
D (Repetitions)	2	34.75	14.18****	12.55****	4.71**

****:$p < .001$ ***: $p < .01$ **: p: $< .025$ *:$p < .05$

1. d.f. for F-ratios were obtained using the Satterthwaite approximation
2. A random; B, C, D fixed effects.
3. A, B, D, random, C fixed.
4. A, B, C, D all random effects.

Is there a significant effect due to "lanes" or to "repetitions?" If these conditions are assumed to be fixed factors (exhaustive sampling; no desire to generalize to additional lanes or more repetitions) the answer is "yes". But if they are assumed to be random samples from larger populations, the answer is "no" since the level of statistical significance has drastically decreased.

Perhaps the choice of ANOVA model is a combination of inputs, from the scientific investigator, and the sponsoring user of the results. From a sponsor's perspective, it may well be that only those conditions which are studied in the experiment are of interest. If many lanes, repetitions, or even tables are never to be studied or added to this testing program, then those factors would never be sampled from a larger population of such factors. However, one might argue from a scientific point of view that many additional lanes, repetitions, and firing positions could have been tested, or will be 'tested' in real life. That is, we happen to have chosen only three repetitions, two lanes per subject and eight different distance-position combinations. Thus, the sponsor-practitioner wishes information that is specific to his particular test. In contrast, the scientific "purist" may perceive this one test or experiment as merely one of many different kinds which could have been conducted by him for the sponsor. Ultimately, the choice of model can influence the significance levels obtained.

For the last topic to be discussed in this paper, we examine a summary statistic which can be reported along with the F-ratio, but which does not involve a probabilistic inference about the likelihood of the results. This is the "magnitude of effect" or "proportion of variance accounted for" index, omega squared (ω^2). Basically, the magnitude of effect (m.e.) measures the degree of association between the independent variable(s) and the dependent variable(s). In the simplest case for ANOVA having fixed factors, none of which are repeated, the m.e. formula is:

magnitude of effect = $(SS_{effect} - df_{effect} \times MS_{error})/(SS_{total} + MS_{error})$.

Rules for deriving m.e. indices are provided by Dodd & Schultz (1973), along with tables for representative ANOVA designs. The greater the proportion of total variance attributable to treatment effects, the more confident can the investigator be that the effects are sizable and important.

The first step in obtaining the m.e. is to compute the expected mean squares for the particular design. This is most easily done using the Cornfield-Tukey algorithm, examples of which are found in Kirk (1968), and Winer (1971). Then, in order to compute the value of the variance components for each source of variance ("effect" or "treatment" condition), note that the estimated value of each EMS is the observed Mean Square from the ANOVA table. The calculations require that the expression for each EMS be equated to the appropriate values of the mean square. The resulting set of equations is then solved; the general form for the equation being:

$$\text{variance of effect} = \frac{(\text{MS effect} - \Sigma \text{ variance components times their number of levels for all other terms in the EMS effect})}{\text{Number of levels in that effect}}$$

These calculations can be simplified if the most complex interaction term is done first, and you then work your way up to the main effect terms. Then, to calculate the proportion of variance for each source, first sum the values for each component to find the total variance. The proportion of variance for each source equals the value of its variance component divided by the total of all the variance.

Let's return now to the M.P. Firearms Qualification Course example, and see what the ω^2 statistics are for the same sources that were listed in Table 1. These results are shown in Table 2, where it may be seen that the largest effect, other than random error, was due to the "Tables" factor, which captured a 24% share of the total score variability. The effect due to Persons, reflecting individual differences among the students, reached 10%. The effect due to Repetitions in Table 1 was statistically significant, ($p<.01$), whereas according to Table 2, Repetitions contributed an effect worth only about .4%. The reason for this apparent discrepancy between the two summary statistics is due to the large number of subjects, which in turn produced a large number of degrees of freedom. This allows small F-ratios to more readily achieve statistical significance. Thus the values for m.e. in Table 2 act as a check upon the significance levels listed in Table 1. Therefore, the effect due to Repetitions reveals a slight, but probably inconsequential learning effect. A similar line of reasoning holds for the interpretation of the Scores variable in Tables 1 and 2.

TABLE 2. Magnitude of Effects for Each Main Effect Under Different Sampling Assumptions

Proportion of Total Variance, Assuming:

Source	A Random B, C, D Fixed	A, B, D Random C Fixed	A, B, C, D Random
A (Subjects)	.0852	.1027	.1030
B (Lanes)	.0004	.0006	.0005
C (Tables)	.1643	.2454	.2631
D (Repetitions)	.0027	.0041	.0042

The discrepancy between the F-ratio and m.e., as summary statistics shows up most clearly in some of the interaction terms. For example, the AC, AD, CD, and ACD terms were all statistically significant at $p<.001$.

Yet the proportion of variance accounted for was relatively small. Table 3 shows these values.

TABLE 3. Statistical Significance and Magnitude of Effect for Some Interaction Terms

Source	P	M.E.	P	M.E.	P	M.E.
AC	.001	.101	.001	.054	.001	.057
AD	untestable	.049	.001	.044	.001	.037
CD	.001	.002	.01	.003	.01	.004
ACD	untestable	.182	.001	.077	.001	.081

The results are at first glance more equivocal than the main effects presented in Table 2. We see that an interaction source can be highly statistically significant ($p<.001$ or $<.01$), and yet account for an obviously insignificant amount of the overall variance (e.g., .2% to .4% in the use of CD, depending upon the model).

Summary and Conclusions

Internal validity refers to the degree to which a training program results in immediate learning or changes in skills, attitudes, behaviors, etc. External validity refers to the degree to which these immediate changes transfer to on-the-job changes. Internal validity is necessary, but not sufficient for external validity. Selection of an experimental design and statistical treatment of the data can have a powerful impact upon the degree of perceived internal validity.

This paper presented several statistical-experimental design methods which can help to assure higher internal validity in studies which purport to measure training effectiveness. An extended control group design will help to control for various nuisance variables, like aptitude and pretest effects. Statistical inference is a tricky business, and the researcher should know that classical hypothesis testing is just one way of reaching a conclusion following data analysis. Erroneous conclusions following even so common a method as ANOVA can be avoided if a well thought-out sampling plan determines the mooel selection, and if the ω^2 index is used to supplement the F-ratio.

References

Bunker, K. A., & Cohen, S. L. The rigors of training evaluation: A discussion and field demonstration. Personnel Psychology, 1977, 30, 525-542.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Carver, R.P. The case against statistical significance testing. Harvard Educational Review, 1978, 48, 378-399.

Cronbach, L. J., & Snow, R.E. Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington, 1977.

Dodd, D. H., & Schultz, R. F. Jr. Computational procedures for estimating magnitude of effect for some analysis of variance studies. Psychological Bulletin, 1973 79, 391-395.

Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart, Winston, 1963.

Kirk, R. E. *Experimental design*: *Procedures for the behavioral sciences*. Belmont, CA: Wadsworth, 1968.

Mahoney, M. J. Experimental methods and outcome evaluation. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 660-672.

Rosenthal, R. *Experimenter effects in behavioral research*. New York: Academic Press, 1969.

Solomon, R. L. An extension of control group design, *Psychological Bulletin*, 1949, *46*, 137-150.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw Hill, 1971.

THE STATE-OF-THE-ART OF COMPUTER-AIDED COUNSELING AND TESTING

Arthur L. Korotkin
Joanne Marshall-Mies

Institute for Behavioral Research
2429 Linden Lane
Silver Spring, Maryland 20910

INTRODUCTION

Computers have become an integral and accepted part of our daily life. In some cases we take them for granted, and in many cases we are unaware that they are even there. Many institutions such as banks, supermarkets, transportation systems and our employers are using computers to store, process, schedule, record transactions, and supply management information in a ready and useful format. The data processing technology of today has made possible a broad expanse of benefits ranging from automatic checkout at the supermarket to such esoteric applications as computer enhanced photography from satelites.

Despite criticisms sometimes made about automation, there are many functions that computers can indeed perform better than humans: tasks such as sorting, retrieving and computations; and applying logical rules for processing information. Despite their superiority in some areas, it is important to keep in mind that the computer is meant to be used to support and enhance human capabilities--not replace them.

One criticism often made is that fact that the computer is impersonal and reduces human beings and their characteristics to a series of alpha numeric characters. While this may be true, it is balanced by the fact that there are advantages to be gained from anonymity and objectivity.

COMPUTER-AIDED COUNSELING

The computer posseses some unique characteristics which lend themselves to the counseling application. Because of its large memory capacity and ability to access and retrieve a vast amount of stored information, the computer offers numerous benefits to both counselee and counselor. First of all, the computer becomes a primary source of information that is always there for all users according to their individual needs, interests, and avenues of inquiry. Besides accessibility, the computer offers any part of its stored information without bias or partiality. The computer can meet the user at his or her own level utilizing a variety of resources including visual displays, visual graphics and audio. The feedback is rapid and can be diagnostic as well as informative. Most importantly, the computer disseminates information in both an untiring and nondiscriminating fashion. The computer can respond equally effectively to a career question whether it has been asked once or a thousand times within the course of an hour. The data provided by the computer are consistently accurate and unchanged regardless of the sex, race, ethnic background or socioeconomic

(1)

status of the individual requesting the information. In addition, when used in the "interactive mode," the user becomes an active participant rather than a passive observer.

Types of Users

Current user organizations can be divided into four general types: educational institutions, education-related institutions, social service organizations, and employment-related agencies. Boundaries between these categories are not always clearly marked. Educational institutions include elementary, junior high and high schools; junior and community colleges; and two-year and four-year colleges and universities. Education-related institutions include technical and vocational schools, training programs, vocational rehabilitation offices, and libraries. Social service organizations cover programs for the elderly, welfare recipients, high school drop-outs, the handicapped, persons in correctional/penal institutions, unemployed youth, etc. And finally, employment-related agencies--the group which easily overlaps social service organizations--include CETA and WIN manpower agencies, youth employment training, EOC's, women's programs, and in a few instances veteran's groups.

Computer-aided career counseling systems can be used directly by the counselee and/or the counselor. The counselee is the person who must make the ultimate decision--it's that person's career which is at stake. Thus, some systems are geared primarily for use by the counselee in an individual, anonymous, private interactive mode. On these systems the counselee alone decides what information to enter into the system and selects the access strategy which he/she will use. Other systems are designed with the career counselor in mind. In this case, the counselor becomes the intermediary between the counselee and the computer data bank.

User Access and Types of Information Files

Two major methods of accessing the information files are used: direct access and structured access. Direct access allows the user to retrieve directly any piece of information in any of the information files without going through a specified routine. For example, the user might request the annual salary of a high school physics teacher, the amount of education required to be a chiropractor, or the nearest educational institution which gives a B.S. degree in nursing. Contrary to such direct access, the structured access strategy guides the user into the information files through a series of pre-established steps. For example, the user might answer a series of questions related to his/her background, interests, educational aspirations, etc. These data in turn would be used to narrow down the range of occupations which would be described. In order to modify the occupation list, the structured access user would need to re-enter responses to the questions.

Computer-aided career counseling systems vary widely as to the number and type of information files available. In general there appears that the information files can be categorized into four groups: occupational information, educational information, training information and other miscellaneous information. Occupational information files all contain descriptions of specific occupations including such factors as job duties, environmental conditions, salary, educational requirements, amount of leisure or travel, etc. Some systems also group occupations into clusters or similar occupational groupings.

Other occupational information files include VISIT or PEOPLE files which contain names of persons in the occupation and/or interviews with such a person, and Career Resource and Bibliographic files which contain suggested readings or information sources.

Educational information files include career preparation data, school and program data, subject data dnd financial aid and scholarship information. Various institutions are described as to location, size, major courses of study, etc. Training files are similar to education files and describe training, vocational and apprenticeship schools and programs. And finally, the last "Other Miscellaneous" files group includes such information as job search skills, help file for getting a job, summer jobs, job bank summaries, employer descriptions and women's career resources.

Accessing strategies, in conjunction with having current information files, are perhaps the key distinguishing characteristics of the currently operating computer-aided career counseling systems. They also appear to be the critical determinants of the amount of time spent on the system by each user and the perceived usefulness of the system.

Application in the Military

Computer-aided career counseling has its roots in the educational environment where the early systems were developed for career guidance in secondary schools. In more recent years, computer-aided career counseling systems have been developed with a larger audience in mind. Individual states have developed multi-audience systems which are being used in educational, training, social service and employment settings. In addition, the military services are in the process of developing systems with potential use by individuals at all stages of their military careers, from pre-accession to retirement.

As one example the specific areas in which the computer can be a help to a recruiter can be identified. Borman, Hough and Dunnette (1976) in their development of a behaviorally-based rating scale for the evaluation of Navy recruiters listed eight major categories of recruiter activities. They are listed below with an asterisk next to ones in which the computer can be an aid.

*1. Locating and contacting qualified prospects.
2. Gaining and maintaining rapport.
*3. Obtaining information from prospects and making good person-Navy fits.
4. Salesmanship skills.
5. Establishing and maintaining good relationships in the community.
*6. Providing knowledgeable and accurate information about the Navy.
*7. Administrative skills.
*8. Supporting other recruiters and the command.

What is important about the computer's contribution is that it frees the recruiter to do those things that only a human can do, i.e., those activities requiring interpersonal interaction and skills.

COMPUTER-AIDED TESTING

One of the most important data banks in any computer-based career counseling and guidance system is the information which exists about the counselee. One major component of this data bank will be test results. In the area of testing the computer can serve many different roles. It can serve merely to store the data as test scores are entered into the data bank; it can be used to score tests using specially designed answer sheets or forms; and finally it can be used to administer a test to a client, record the answers, score the test and store the data for future reference.

Computer-aided testing has its roots in the educational community where Pressey sought a technique for providing more immediate feedback from tests to the student. He began examining better ways to administer and score objective (multiple-choice) tests. The earliest Pressey machine was a self-scoring multiple-choice apparatus which was exhibited at the American Psychological Association meetings in 1924 (Lumsdaine and Glaser, 1960). It was a mechanical device which displayed the question and answer choices in a small window and provided four keys to input the answer. The key press registered the response and moved the test to the next question.

Several variations of a self-testing device were developed in the next few years However, the real major impetus to self-scoring testing came with World War II when War Department interest led to the development of both improved mechanical devices and refinement of the punchboard apparatus to facilitate military testing and scoring.

In 1944, what may be considered the first large scale digital computer, the Harvard-IBM Mark I, was completed. About the same time as the electromechanical Mark I was being built, the first electronic digital computer was being developed. It was completed in 1946 at the Moore School of the University of Pennsylvania. It was called the Electronic Numerical Integrator and Computer or ENIAC (Baker, 1975). From the punchboards, mechanical devices, chemical devices, and electromechanical counting machines, teaching and testing moved into the electronic age and to electronic computers.

In the early stages of applying automation to testing the applications were usually limited to the automatic scoring of paper and pencil tests. Such application still exists and most major universities and colleges have test scoring machines which could be used to scores examinations developed by the instructional staff and some standardized tests administered at or by that institution. In addition, the major test developers and testing firms maintain automated scoring services for some of the instruments they publish, which may be too complex or time consuming to score, or require regional or national scoring in order to arrive at normative data.

In discussing the application of technology to testing, Hieronymus (1972) pointed out some advantages of the computer.

- Machine-scorable booklets have opened up the possibility of using certain types of items not previously practiced.
- Can get complete, complex, and inexpensive tryout data.

- Can get more efficient standardization and have complex and multiple types of norms prepared.
- Allows for complex analysis of the stored data.
- Can assemble and present tests by computer from stored item banks.

This latter item begins to expand the role of the computer in testing and opens up the realization that beyond automated test scoring lies several levels of increased sophistication in computer-aided testing, such as:

- Computer administered, scoring and interpretation of tests.
- Truly interactive systems wehre computer administered and scored data are combined with other data on that individual already in the data bank. Additional testing is dependent upon the results of these analyses which may be done at a later date or "real-time" while the individual is still seated at the test station.
- Adaptive or "tailored" testing where the responses of the individual taking the test determine subsequent test items.

Computer Administered Testing

The next step to having the computer score the test is to have it administer the test. Carroll (1971) mentions this possibility in discussing computers in testing. In listing the advantages of the computer he went further than mere scoring:

- Since it stores and analyzes responses, the computer can easily perform item analyses.
- Standardized tests can easily be administered and scored, and then produce the information in desired format.
- "Use of consoles in remote location might make possible the administration of standardized tests simultaneously over wide geographical areas--even computerized nationwide test administration (as of College Board tests) is not out of the question" (Carroll, 1971, p. 820).

At first the paper-and-pencil tests were merely automated by using the computer to administer, score and interpret the tests. Some of the first such efforts were in the area of personality tests. Later, such automation was used with intelligence and ability tests. Kleinmuntz (1975) has pointed out, in his discussion of "The Computer as a Clinician," that the computer can not only compute but it can perform other "noncomputational tasks." It can read, translate, compare and associate. Therefore, it can be used to do "configural item scoring," that is, scoring for patterns of items rather than individual items on a test such as the MMPI (Williams and Kleinmuntz, 1969).

Carrying "profile analysis" further, Kleinmuntz (1975) suggests that computers can go on and be programmed to interact with people by creating interactive interview dialogues. Clearly as the testing procedure includes the gathering of other data we begin to go beyond computer test administration and into Interactive Testing Systems.

Interactive Testing Systems

Truly interactive systems are fairly recent innovations. Aside from the counseling and guidance systems, which are the major focus of this report, most of the other interactive systems are found in a clinical setting. To the extent that the clinical diagnoses and admitting strategies are in fact similar to assessment and job placement strategies, it may well be appropriate to discuss a prototype computer-asisted admissions system as an existing example on Interactive Testing Systems.

The Computer-assisted Psychiatric Assessment Unit (PAU) was developed to optimize the assignment of patients into the treatment system in order to improve individual care and the maximum utilization of existing facilities (Johnson and Williams, 1975). It was established at the Salt Lake City Veterans Administration Hospital in 1973.

By using an on-line computer, psychological, social and physical assessments are made at the time of application for care. The computer summarizes, interprets and prints reports on the patients so that the data are available for immediate decision making. A large amount of information is collected "on-line" using a cathode ray tube (CRT) terminal. The procedure begins with entering basic identifying information into the computer for a new patient. If the patient is capable of being tested the PAU coordinator administers a mental status examination and enters the information into the terminal. (A copy is printed remotely in the PAU office area.)

The patient is then given instructions for completing the self-reporting testing. The battery includes a medical history, several personality tests and an Intellectual Performance Test. A health technician performs a computer-prompted medical screening physical examination. These results are also entered into the computer via the CRT. The data are analyzed and reports prepared by computer. These are reviewed by the PAU staff and some form of treatment recommended.

Prior to PAU, patients were subjected to a lengthy evaluation process before treatment could begin. After almost 6000 cases, data indicate that staff time devoted to intake evaluation has been reduced from 13.5% to 4%. In addition, there is evidence to suggest that there has been a positive effect on patient care.

Computer Adaptive Testing

Computer Adaptive Testing (CAT) is the use of the computer to dynamically select test items for presentation based on the individual's responses to previous items. The test is adapted or "tailored" to the unique capabilities, knowledge or achievement of the respondent by selecting the appropriate items from a large item pool stored in the computer. The computer selects and sequences these items for the person being tested in a similar manner to the way instructional materials are chosen and presented in computer-assisted-instruction (CAI). The key value of adaptive testing lies in the fact that most tests are designed for a broad spectrum of individuals so the test must discriminate over a large range of difficulty levels. This means that there are not many discriminating items at each level of achievement of proficiency. Using CAT, it will be possible to concentrate the test items at the appropriate level for each individual.

The advantages of adaptive testing have been described by Weiss (1976), McBride (1977) and Steinheiser (1979). These are summarized below:

- Reduction in testing time.
- More precise measurement over a wider range using fewer items than conventional tests.
- Reduction in variance caused by frustration, boredom and test anxiety.
- Reduction in human errors in marking answers, test scoring and recording of the results.
- Reduction of the chances of test compromise by eliminating test booklets and by the fact that the tests are different for different individuals.
- Improved diagnostic or profile information.
- Rapid feedback of test results.
- Expected reduction in classification errors of both types, i.e., rejecting people that shouldn't be and accepting people that should be.
- Good practical potential for application to military manpower testing, scoring, selection, classification and job counseling.

The utility of adaptive testing to military applications has been recognized for some time, and some of the pioneering work was done in military laboratories. Its importance is underscored by the recent formation (June, 1979) of the Computer Adaptive Testing Joint Service Coordination Committee, a Department of Defense effort with the Navy designated as a lead service, to examine the use of CAT for accession testing within the next 4-5 years.

CONCLUSIONS

In a military setting a comprehensive computer-aided counseling and testing system has the potential for performing the following functions:

(1) Record biographical and background data on individuals.
(2) Administer, score and interpret tests.
(3) Store information on counselees.
(4) Store information on available job and school openings.
(5) Aid in job and school assignments.
(6) Provide occupational and educational information to counselees and counselors from the time the individual enters the service until retirement.
(7) Provide accurate record keeping, data analysis and the production of administrative reports.

REFERENCES

Baker, J.C. The computer in the School. Bloomington, Ind.: The Phi Delta Kappa Educational Foundation, 1975.

Borman, W.C., Hough, L.M., & Dunnette, M.D. Development of behaviorally based rating scales for evaluating the performance of U.S. Navy recruiters (NPRDC TR Rep. 76-31). San Diego, CA: Navy Personnel Research and Development Center, February 1976.

Carroll, J.B. In Tickton, S.G. (Ed.), To improve learning: An evaluation of instruction technology. New York: R.R. Bowker Co., 1971.

Hieronymus, A.N. Today's testing: What do we know how to do? In Educational Testing Service. Proceedings of the 1971 invitational conference on testing problems. Princeton, N.J., 1972.

Johnson, J.H., & Williams, T.A. The use of on-line computer technology in a mental health admitting system. American Psychologist, March 1975, 30 (3), 388-390.

Kleinmuntz, B. The computer as clinician. American Psychologist, March 1975, 30 (3), 379-387.

Lumsdaine, A.A., & Glaser, R. (Ed.). Teaching machines and programmed learning: A source book. Washington, D.C.: National Education Association, 1960.

McBride, J.R. Adaptive mental testing: The state of the art. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, Draft, March 29, 1977.

Steinheiser, F. Symposium: Military application of computerized adaptive testing. Proceedings of the 87th Annual Convention of the American Psychological Association, 1979.

Weiss, D.J. Adaptive testing research at Minnesota: Overview, recent results, and future directions. Proceedings of the first conference on computerized adaptive testing, March 1976.

Williams, J.G., & Kleinmuntz, B. A process of delecting correlations between dichotomous variables. In Kleinmuntz (Ed.), Clinical information processing by computer. New York: Holt, Rinehart & Winston, 1969.

COMPUTER-AIDED CAREER INFORMATION SYSTEMS

Bertha H. Cory

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

INTRODUCTION

Computer-aided career development research at the Army Research Institute (ARI) to date has focused on the Army officer, from the point of his entry into service up to the grade of Colonel. As requested originally by the Deputy Chief of Staff for Personnel, ARI's efforts have been designed to recognize (1) objections expressed in various surveys of junior officers regarding existing career counseling (findings of insufficient career information, inadequate help at decision points, sources not up to data or not easily available, seeming ignoring of preferences, difficult or limited access to career counselors); (2) requirements of the Officer Personnel Management System (OPMS); (3) current theory and research in career counseling psychology; (4) recent computer technology as successfully applied to career counseling in education and industry. ARI research has designed the Officer Career Information and Planning System (OCIPS), which will be described here.

Also briefly described here is a new initiative involving design of a computer-aided information system for the Army Continuing Education System at Army Education Centers.

BACKGROUND

The recent identification of career planning as a salient issue for adults and the scarcity of professional assistance in that problem area has led career development specialists to seek alternative approaches to educational and career planning. One such approach is the computer-aided guidance system. A number of such systems have been developed in a variety of settings, primarily intended for use among high school and college students. OCIPS is one of the first designed to enhance the career planning skills of adults whose careers are already in progress.

To provide assistance in the problem area of long-term educational and career planning, career development specialists have sought to adapt available resources in computer technology to specific career planning tasks. Current approaches to the use of computers in educational and vocational guidance are based on two theories. First is the pragmatic theory which asserts that the more information that is available to individuals, regarding both self and world of work, the better their vocational decisions are likely to be. Second is the developmental decision-making theory which reflects the view that a career develops--and, thus, decisions are required--over the life span, rather than as a result of specific, point-in-time, educational or vocational choices.

The technological capacity of the computer is widely recognized and has considerable potential when applied to the process of career decision making. The computer provides the capacity (1) to store, retrieve and update large amounts of data, (2) to interrelate data about individuals and their environments, (3) to individualize data to generate educational and career alternatives, (4) to simulate conversations of interviews through interactive terminal devices, (5) to modify user behavio to provide feedback, review and personalized assistance to counselor or client, (6) to control and coordinate audio and visual material with text, and (7) to provide services to many users simultaneously in various settings. By making use of these capabilities, career development specialists have been able to provide the needed assistance to individuals at their different developmental stages in terms of information gathering, vocational exploration, and career decision making.

OFFICER CAREER INFORMATION AND PLANNING SYSTEM (OCIPS)

Drawing on theory and research in counseling psychology and technologies in computer science, OCIPS is envisioned as a computer-aided career information and planning system for Army officers. This system can provide a number of benefits to the Army officer and to Army management, including:

--greater ability of an officer to take responsibility for his or her own career decision making;

--greater officer satisfaction and increased knowledge of the career-enhancing potentialities of various assignments;

--better fit of officer to job based on the consideration of aptitudes, values, interests, education, training, and experiences; and

--greater equity and efficiency in the career management system.

In order to begin to accomplish these goals, the initial phase of the system's development called for a long-range career planning dialogue unit that would enable Army officers to explore planning strategies and decision-making techniques and to develop and apply career goals and values to their own long-term career planning. It was decided that the system would need to conform to a number of specifications. First, the dialogue units should allow the officer to explore career-related values and strategies for implementing those values. The units should advocate flexibility in career planning and be applicable to Army careers. Second, the dialogues should appear as a natural conversation between an officer and a human counselor, using explicit, concise language tailored to Army officer background and interests. Finally, the dialogues should be designed to increase the officer's awareness of the notion of a career as a time-ordered sequence of positions, mediated in part by his or her own choices, in the milieu of the requirements of the Army.

Conceptual Development. From these objectives and specifications a set of computer-aided experiences for teaching various career planning concepts and for enhancing career-relevant competencies was created. The specific concepts, on which the long-term career planning portion of this system is based, are those that emerged from Super's (1957) longitudinal study of career development. The concepts represent those notions that research has shown to be essential for consideration in career planning. They are:

The inevitability of choice: stressing the opportunity and obligation on the part of an individual to make certain choices, and reviewing the consequences of not choosing when choice is indicated;

Choice as an implementation of values: introducing the notion that the major determinant of any given choice ought to be the values of the chooser, necessitating some clarity about one's own value system;

Contingencies and discontinuities: making explicit the implicitly obvious existence and influence of events that one is unable to predict;

Clarity and tentativeness: illustrating the necessity of having clear, well-designed plans, while simultaneously recognizing the unavoidable tentative nature of such plans; and

Life stages: focusing on the available knowledge of career-related behavior at different points in an individual's development.

Such research has also shown that certain competencies in career planning tasks must be developed in order to negotiate a career successfully. Drawing on this body of knowledge, the system was to include exercises in the following areas:

Skills and values clarification. An officer's skills and values are major determinants of career satisfaction and success. Accordingly, the ability to identify these primary skills and values is important to career development. This ability has other important components: (a) recognizing and resolving conflicts among values and skills; (b) recognizing the linkage between specific values and skills and career decisions; and (c) preparing for possible revisions of the primary skills and values throughout one's career.

Career strategies. Career planning requires the ability to translate self-and-environment knowledge into planful action. The components of this are: (a) interpreting life goals in light of one's primary values and skills; (b) developing life goals that are optimally enhancing for career development; (c) harmonizing conflicting goals; and (d) developing action plans for reaching specific objectives. Overall, "career strategies" means the ability to implement one's primary values and skills in specific, concrete actions.

Choice point identification. In a complex career system, it is important to be able tc identify those points where one can choose and where that choice can make a significant difference. This includes the ability to anticipate future choices, to evaluate present choice alternatives, and to assess the irreversibility of specific choices.

Career monitoring. Assessing career progress is important in view of the tentative nature of career planning. Assessment requires a systematic way of continually integrating the career environment with one's primary values and skills.

System Description. The system developed thus far consists of seven interactive, or conversational, dialogue units. The user's path through the units is determined by his or her several choice points within each unit. Each module is self-contained and connects with the other modules via an executive monitoring system. To date, SIGNON, FORESIGHT, OVERVIEW, and ALTERNATE SPECIALTY have been programmed on the UNIVAC 1108 and are useable from terminals in demonstration form. The remaining modules--CAPTAIN'S INTRODUCTION, SELF-ASSESSMENT, and CAREER STRATEGIES--are in script form and have nct yet been programmed. The various modules are described below.

SIGNON. This introductory module introduces the officer to the system, instructs the officer how to use the terminal, and asks for a variety of identifying data such as military specialty, type and level of civilian education, and current military status.

FORESIGHT. This module is designed to introduce the user to long-term career planning. It begins with consideration of the belief that individuals can influence their career progress if (a) they know what they want, and (b) they know how the system works. The basic career concepts described earlier are assigned code names "Must"--choice is inevitable; "Value"--you have to know what you want; "Surprise"--unexpected events happen even if you plan; "Tension"--simultaneously firm and tentative planning; and "Stage"--predictable life changes. The user may elect to look through any or all of the five- or six-frame interactive explanatory illustrations for each concept. The conclusion of the module integrates the concepts in a sample career path that shows an officer making choices and confronting situational changes at different stages in his career. The ability to convey to the user the most current available knowledge about career planning and career development in an understandable and thought-provoking manner is the most outstanding quality of the FORESIGHT module.

OVERVIEW. This informational module includes the Army's overall plan for the progression of an officer's career and attempts to make the user aware of those factors which can influence the way in which an officer's career develops. Those include:

--changes in needs, goals, and objectives of the Army
--military and technological changes
--timing of career decisions
--Officer Evaluation Reports
--military education
--specialty assignments
--civilian education and training

It dissects the patterns and determinants of Army careers with the use of a series of off-line charts and offers the user answers to a series of typically-asked questions. It reinforces the concepts introduced in FORESIGHT and adds some Army-specific concepts such as officer responsibility and dimensions of utilization and training. OVERVIEW facilitates the officer's comprehension of "how the system works"--a necessary ingredient in career decision making--and does so in a manner that enables officers to incorporate the understanding of the complex officer career progression system into their planning.

CAPTAIN'S INTRODUCTION. Experience with the system has shown that, while younger officers (lieutenants) profit from FORESIGHT and OVERVIEW, officers who have achieved the rank of captain or above have already acquired much of the information contained in the modules. Therefore a substitute introductory module was designed for users already familiar with the army career progression system. This module, called CAPTAIN'S INTRODUCTION, includes the information in FORESIGHT and OVERVIEW in a more abbreviated form.

ALTERNATE SPECIALTY. One of the system's long-range objectives is to provide the user with access to real data relevant to important choice points in an Army officer's career. The submodule of OVERVIEW and CAREER STRATEGIES, called ALTERNATE SPECIALTY, is an example of how this can be done. Due to the implementation of dual occupational specialties for Army officers, expressing a preference for an alternate specialty is a critical choice point in an officer's career. A rich data file relating officer characteristics and preferences to alternate specialty designation affords the user a unique opportunity to engage in meaningful career exploration. The ALTERNATE SPECIALTY submodule was developed to make use of this data file and includes information about the alternate specialties that are available, how they are designated, and how career plans can influence them. In making the data available to the user and in offering suggestions about useful ways to interpret them, the submodule provides the officer with the opportunity to explore and compare his or her characteristics with those of officers for whom any given specialty was designated during the previous year and to integrate this information into an effective career strategy. (Note: While retaining the concept of dual specialties, the Army has recently ceased to refer to them as "primary" and "alternate"; this does not affect the utility of the module.)

SELF-ASSESSMENT. Other modules (OVERVIEW and ALTERNATE SPECIALTY) have addressed the issue of "how the system works." The SELF-ASSESSMENT module is designed to help users clarify "what they want"--a necessary component of satisfactory career planning. The officer uses a representative list of skills and values to create an individualized profile based on preference and performance (skills), and subjective importance (values). This list of skills was derived from an analysis of Army officer job performance dimensions and available inventories of relevant career skills. Similarly, the values list represents a combination of work value inventories, lists of values used in industrial personnel development programs. and values derived from ARI surveys. Once the officer has created a profile, suggestions are offered about integrating self-assessment into planning, and the user is asked to evaluate previous and anticipated assignments in light of this profile.

CAREER STRATEGIES. This module is designed to help officers implement their career aspirations through exercises in setting long-term goals and in translating goals into action plans for immediate objectives. The introduction conveys to the officer:

--that goals provide the basis for long-term planning;
--that goals are arrived at by assessing the structure of Army career opportunities and by assessing one's own characteristics;
--that long-term goals can only be obtained by achieving intermediate objectives; and
--that concrete plans of achieving intermediate objectives provide the link between career planning and intelligent action.

The process of creating a career strategy is introduced by the use of a career planning game ("SCOR") which incorporates the major aspects of an officer's career: military specialties, education and training, skills, job performance, rank, assignments, family, and values. The game uses an off-line playing board ("SCOR-BOARD") for charting hypothetical career progression. The decision points in the game require the player to deal with four career issues: the inevitability of Surprise, the necessity of Choice, the awareness of Opportunities, and knowledge of Requirements. The player starts the game as a second lieutenant, selects pre-programmed goals, seeks to move toward those goals in a series of computer managed decisions, and arrives at an end point that signifies goal achievement.

At the conclusion of the game, the principles of creating career strategies are reviewed and the user is presented with the "Career Planning Wheel." This off-line chart is similar to the SCOR-BOARD, but depicts the major aspects of an officer's career in more detail. The user may access computer-based career data related to the year of commissioned service in each aspect of the wheel.

After the SCOR game and the Career Planning Wheel have illustrated the use of career strategies and career information, the user is asked to review his or her own career goals. Each goal is examined with a series of eight criteria for effective career planning goals and is revised until it satisfies the criteria. The revised goals are then translated into action plans for intermediate objectives. For example, users are guided to convert goals to actions by choosing a specific standard for gauging success, identifying resources and barriers, setting checkpoints and deadlines, and so on. The results of this module include clarified career goals, intermediate objectives, and action plans which have been tested for their adequacy.

System Evaluation. In order to assess potential operating difficulties and to obtain some initial reactions to the acceptability and usefulness to the target population, four of the seven modules (SIGNON, FORESIGHT, OVERVIEW, and ALTERNATE SPECIALTY) were field-tested on 52 company grade officers at Fort Benning, Georgia. Each officer was administered a pre- and post-use instrument measuring: knowledge of relevant information; attitude toward the computer as a guidance tool; and the understandability, accuracy and usefulness of each module. A post-use debriefing session was also conducted.

The results of the field trial were extremely encouraging. The users found the content of the modules to be interecting, accurate, useful, and understandable and gave highly favorable ratings to the use of the computer as a method of transmitting career information. Those officers who used the system reported a decreased need for career information and an increased level of certainty and satisfaction with alternate specialty preference. It was also found that the style and the humor of the text were considered appropriate and enjoyable.

Although no major revisions were indicated by the field trial, several adaptations and expansions of the system have since been suggested. Designed primarily for the use of company grade officers (those who have fewer than eight years commissioned service), OCIPS could well be adapted for use by field grade officers, including particularly those who had used the system in their years as junior officers. Other potential users are those who are involved in officer career management. Alternative entry modules, similar to the CAPTAIN's INTRODUCTION, would introduce such personnel to the system. Proposed content expansions include a module dealing with the career planning and decision making necessary upon severance from military service. Expanded data bases could add information about military and civilian education and extended longitudinal data relating alternate specialty designation and other career events to later career plans. The system is also capable of administering and scoring assessment instruments and could be equipped with the capacity to monitor and store patterns of system use for research and re-evaluation.

OCIPS currently exists in the first generation phase of development. Those components which have been tried in the field work well and are acceptable to the target population. CAREER STRATEGIES requires additional data and dialogues for full operation. Also needed are additional data banks and further development of career monitoring. However, OCIPS has demonstrated potential for expanded operations. Further field testing and subsequent revision would be appropriate before operational implementation.

In terms of ARI's programs of research funding, OCIPS can be viewed as having completed its "basic research" and "exploratory development" ("6.1" and "6.2" monies) and being ready for the next phase of "advanced development" ("6.3" monies). Moving into the next phase has been temporarily deferred, awaiting (1) a predictable change in the milieu of computer technology on Army posts--more specifically the commonplace presence of CAI/CMI capability upon which the career counseling function can be readily "piggy-backed" without significant additional cost; and (2) resolution of philosophical differences within Army management--to what extent the junior officer should be encouraged to plan and develop his personal goals and interests--or should the junior officer be indoctrinated to consider only the Army needs--that is, what he can do for the Army rather than what the Army can do for him.

ARI has career counseling modules within the structure of OCIPS that are available for interactive demonstration using the UNIVAC 1108. These modules go far beyond simple "page turning" and are innovative with respect to both counseling theory and application of computer technology. Reports are available; a bibliography relating to ARI products in this area can be requested.

ROTC CAREER COUNSELING SYSTEM

Another career research effort which will be on-going as a new initiative in the coming year is design of a computer-based career information and planning system for ROTC cadets. Content and methodology from OCIPS will be used and/or adapted as appropriate.

ARMY EDUCATION INFORMATION SYSTEM

In the spring of 1979 research was initiated on design of a system for interactive, computer-based delivery of information on the Army Continuing Education System (ACES). ACES is a broad program of largely voluntary self-development in many different areas--both academic and non-academic--with a goal of providing to each soldier educational opportunities at least equal to those available had he or she not entered on active duty. The system being designed would be installed via terminals at each Education Center, would allow soldiers to obtain general as well as specific information about the many ACES opportunities, and would allow exploration of options before conferring with an educational counselor. The system--which could also be accessed by the counselor--would relieve counselors of clerical work, keep information up-to-date, retrieve local information from other posts, relieve counselors of much repetitive routine information-giving, and generally allow counselors more time to concentrate on guidance activities. Furthermore, since the system will inevitably include information on progression within the Army enlisted career system, it will provide a significant base for future development of an Enlisted Career Information and Planning System.

REFERENCES

Super, D. The psychology of careers. New York: Harper & Row, 1957.

Super, D., Crites, J., Hummel, R. Moser, H., Overstreet, P., & Warnath, C. Vocational development: A framework for research. New York: Bureau of Publications, Teachers College, Columbia University, 1957.

COMPUTERIZED VOCATIONAL COUNSELING USING THE VOCATIONAL INTEREST-CAREER EXAMINATION (VOICE)

Thomas W. Watson

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

INTRODUCTION

Presently, most psychological testing is administered using a paper-and-pencil format. Computer technology has been advancing at a rapid rate and can be readily applied to test administration and scoring. In fact, during the next few decades, it is probable that most, if not all, testing will be conducted via computer. In order to prepare for the future, the Air Force Human Resources Laboratory (AFHRL) has initiated research to apply such technology to testing and vocational counseling. Other DoD agencies such as the Navy have also initiated similar efforts (Yellen & Foley, 1978). This paper describes one such effort, aimed at development of an automated approach to vocational counseling using the Vocational Interest-Career Examination (VOICE). The VOICE is a general purpose interest inventory designed primarily for use with Air Force recruits to improve job-placement and cross-training decisions (Alley, Berberich, & Wilbourn, 1977; Alley, Wilbourn, & Berberich, 1976). It is currently in paper-and-pencil format and is expected to be implemented in the near future.

Two concurrent approaches have been selected in preparing the VOICE for operational implementation. The first is to develop a simplified hand scoring procedure for use with the paper-and-pencil instrument (Watson, Alley, & Southern, 1979). The second, which is the focus of this paper, is to develop a completely computerized vocational counseling system.

A computerized vocational counseling system based on the VOICE would provide numerous advantages over a paper-and-pencil, hand-scored approach. Some of these advantages are outlined briefly below. Possible advantages are many and the list is not meant to be exhaustive.

1. Scoring complexity would not be a problem due to the superior data processing capability of the computer. It would not be necessary to focus on a simplified scoring procedure amenable to rapid hand-scoring, or scoring by relatively unsophisticated equipment, and a variety of scoring methods could be kept readily available. Optical scanning of paper-and-pencil, machine-readable answer sheets, would not be necessary, saving time and money. In addition, delayed batch processing would not be involved. Overall, computer-driven scoring methods are anticipated to be better suited for the operational environment and superior, from a psychometric viewpoint, to the hand-scoring method also being developed (Watson, Alley, & Southern, 1979).

2. Since scoring is both rapid and immediate, no delay would be incurred and feedback to the user would not be delayed. Also, the format of feedback could be varied. Format could be graphic, numeric, or written, and provided via a cathode-ray tube (CRT) or in hard copy. Feedback could be directed either to other automated personnel data systems, the respondent, or to staff members of user organizations.

3. The time and effort that would have otherwise been devoted to paper-and-pencil administration by staff members in user organizations could instead be devoted to other job tasks since automated procedures could operate with limited staff involvement. Limited staff participation would also reduce human error or bias, and the need for training of staff members in testing or counseling techniques.

4. Administration time could be reduced since computer administration is typically faster than paper-and-pencil administration, and can be paced, if necessary. The current VOICE (Form C) has been reduced in length to 245 items with a mean administration time of 23 minutes with 98% completion in 35 minutes. However, for operational use, it would be advantageous to reduce this time further. If an adaptive scoring method is developed, testing time could be reduced further since only a limited number of items, tailored to individual interests, would be administered.

5. Integration of more information into the counseling process would be facilitated since VOICE data could be integrated easily with information from other automated personnel data systems.

DEVELOPMENT OF COMPUTERIZED VOCATIONAL COUNSELING PROCEDURES

Background

In order to prepare the VOICE for possible use in a computer-adminstered format, AFHRL recently initiated a contract with Psych Systems, Inc., of Baltimore, Maryland, to develop a prototype computerized vocational counseling system. In addition to the development of computerized VOICE administration and scoring procedures, development of an adaptive testing methodology is also an important goal. Since the contract has only recently been initiated, our intentions rather than our accomplishments will be described in this paper. This project began in June 1979 and is expected to take approximately 15 months to complete.

Specifically, this effort is directed toward developing a computerized system capable of administering and scoring the VOICE and providing rapid graphic and numeric feedback to the user, via a CRT and hard copy. The system will also be designed to take other information into account such as aptitude, biographic/demographic, and personality data, and will ultimately be integrated with other automated personnel data systems such as the Automated Personnel Data System/Procurement Management Information System (APDS/PROMIS) (Ward, Haney, Hendrix, & Pina, 1978) and the computerized adaptive aptitude testing (CAT) system (Ree, 1978, 1979). The VOICE computerized prototype is to be designed for use either independently or in conjunction with these other systems.

Specific tasks which are currently being accomplished, or will be accomplished, are as follows:

1. The contractor is currently in the process of developing a comprehensive plan, incorporating alternative options, for implementing such a system. This comprehensive plan is intended to provide a broad overview of possible design options, keeping long-range operational goals in mind. Consideration will be given to such topics as ultimate system configuration, use with various target populations, feedback of information, mobility of system hardware, and types and locations of processors. The plan is meant to provide a basis for informed decisions concerning future courses of action in developing and implementing the system.

2. The contractor has selected the hardware and is developing software necessary for computer-driven administration, scoring, and feedback of VOICE data. For economic and technical reasons, the decision was made to use components of an Automated Measurement System currently in the AFHRL inventory of computer equipment. This system uses two PDP-11/34 processors as hosts for a variety of test stations.

3. The contractor has been investigating, and will compare and contrast, various approaches to adaptive testing, both in the area of aptitude testing, where adaptive methods have been applied, and in the affective domain, such as interest and personality testing, where application of adaptive methods is relatively new. There is a moderate to high degree of risk associated with this enterprise since much of the adaptive testing technology developed for aptitude measurement may not apply to the interest domain. For example, the content of items is vastly different in the two areas, as are scoring techniques. Concepts of item difficulty, item characteristic curves, and right versus wrong answers, do not apply in interest testing. Therefore, approaches to adaptive interest testing, if successful, might be radically different from approaches to adaptive aptitude testing which have already been developed.

4. Once a tentative approach to adaptive interest testing has been identified, the contractor will develop and implement a plan for evaluating the comparative psychometric properties of the computer-driven adaptive method, the paper-and-pencil method using conventional scoring techniques, and the non-adaptive computer-driven method using conventional scoring techniques. This process will involve refinement of the selected adaptive strategy using data from VOICE data files, and administration of the VOICE to a select sample of Basic Military Training (BMT) students in both traditional (paper-and-pencil) and computer-driven (regular and adaptive) formats.

5. Once the psychometric properties of the three modes of administration have been compared, the contractor will develop and demonstrate a functional computerized prototype vocational counseling system capable of administrating, scoring, and providing feedback of VOICE data. The prototype will be able to score the VOICE using existing scoring methods and the adaptive method, and be capable of integrating information from other sources.

6. Finally, the contractor will document all previous tasks involved in prototype development by writing a formal technical report.

IMPLICATIONS

Computerized interest assessment based on the VOICE will have far-reaching implications for VOICE implementation.

Paper-and-pencil VOICE administration with either hand-scoring or remote machine-assisted batch scoring is likely to be the format used when the instrument is initially introduced in an operational setting. In fact, plans are currently being developed for limited use of the VOICE, in paper-and-pencil form, for counseling reenlistment-eligible personnel about possible cross-training opportunities.

Use of the VOICE on a large scale basis may be impeded by its current paper-and-pencil format. A primary use of the instrument is expected to be as a component of the initial job-placement-decision process. Used for such a purpose, it would be administered to large numbers of recruits during processing at Armed Forces Examining and Entrance Stations (AFEES). In these settings, paper-and-pencil administration and hand-scoring or delayed batch processing may not be sufficiently responsive to operational time requirements.

The availability of computer-driven methods of administration and scoring would greatly facilitate large-scale VOICE implementation by allowing the instrument to be more responsive to operational needs, particularly with regard to administration time and ease of scoring. Reduced administration time and scoring ease will be benefits derived from the prototype regardless of the success or failure of efforts to develop an effective adaptive VOICE testing strategy. Other benefits include immediate user feedback and reduced involvement of staff members in the testing and counseling process.

An ultimate goal of development of the computerized VOICE system is integration with other systems into a multi-purpose classification system. Computerized adaptive aptitude testing (Ree, 1978, 1979) is still experimental, and it is expected that ultimately the same equipment will be used in the operational environment for both automated interest and aptitude testing. In fact, the current computerized job assignment system (Ward, Haney, Hendrix, & Pina, 1978) might also be ultimately integrated with these other micro-systems into one macro-system.

Implementation of the computerized counseling system would represent a considerable investment, especially if several CRTs are required at each of several AFEES and perhaps at other locations. However, integration of the automated VOICE with other systems will greatly reduce costs associated with these multiple functions. Also, the computer industry, due to mass production techniques and rapid technological advancement, is one of the few industries in which costs are going down in an inflationary era.

A computerized counseling system based on the VOICE will greatly enhance the ability of the Air Force to make rapid, effective decisions concerning the initial job placement of large numbers of recruits. It will also improve the ability of the Air Force to assist dissatisfied job incumbents in making cross-training decisions. The computer- and interest-assessment technology under development may also be transferable to other settings in government, civilian industry, and education.

REFERENCES

Alley, W.E., Berberich, G.L., & Wilbourn, J.M. Development of factor-referenced subscales for the Vocational Interest-Career Examination. AFHRL-TR-76-88, AD-A046 064. Brooks AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, June 1977.

Alley, W.E., Wilbourn, J.M., & Berberich, G.L. Relationships between performance on the Vocational Interest-Career Examination and reported job satisfaction. AFHRL-TR-76-89, AD-A040 754. Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1976.

Ree, M.J. Implementation of a model adaptive testing system at an Armed Forces Entrance and Examination Station. In D.J. Weiss (Ed.) Proceedings of the 1977 Computerized Adaptive Testing Conference. Minnesota: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1978.

Ree, M.J. An Automated Classification System. Paper presented at the 1978 annual conference of the American Psychology Association, Division 19 Symposium, New York, September 1979.

Ward, J.H., Jr., Haney, D.L., Hendrix, W.H., & Pina, M., Jr. Assignment procedures in the Air Force Procurement Management Information System. AFHRL-TR-78-30, AD-A056 531. Brooks AFB TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, July 1978.

Watson, T.W., Alley, W.E., & Southern, M.E. Initial development of operational composites for the Vocational Interest-Career Examination. Paper presented at the 21st annual conference of the Military Testing Association, San Diego, October 1979.

Yellen, T.M.I., & Foley, P.P. Navy Vocational Information System. NPRDC TR 78-22. San Diego CA: Navy Personnel & Development Center, June 1978.

CAREER GUIDANCE FOR MORE EFFECTIVE PLACEMENT

Dean Halstead

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

Navy Career Counseling, as outlined by Meshi, Holoter, Dow, and Grace (1972), can be divided into three segments, Pre-Assignment Contacts (PAC), In-Service Counseling Actions (ISCA), and Post-Service Counseling (PSC). The PAC segment includes all discussions with recruiters and classifiers prior to an enlistee's reporting to his first post-boot camp duty station; ISCA includes most interviews with the service member between time of reporting to his first duty station and time of discharge; and the PSC discussion covers post-discharge alternatives for the individual who has decided not to re-enlist.

The primary mission of Navy Career Counseling is to increase the Navy's retention rate in order to maintain adequate numbers of qualified personnel in the regular Navy (Meshi et al.). Accordingly, the ISCA segment, which is designed to encourage personnel who are performing satisfactorily to reenlist, has been developed to a greater extent than either the PAC or PSC segments, despite studies which have shown that practices at the PAC stage directly affect retention rates and attrition.

Statement of Problem

Currently, Pre-Assignment Contacts consist of three interviews between Navy personnel and applicants/recruits prior to the assignment of their first duty station or training school. They are the Recruiting Interview (Recruiting Station), Classification Interview (Armed Forces Entrance and Examination Station), and Classification Interview (Recruit Training Center).

The main problem of the PAC segment, as seen by this research effort, lies within the recruiting interview. There are three main areas of this interview in which procedural improvements could enhance the recruiting process: (1) providing more complete Navy career information at the Recruiting Station, (2) ensuring consistency in the information dispensed, and (3) improving the accuracy of that information.

For two reasons, it has been difficult to deal with the first problem, that of providing more complete Navy career information at the Recruiting Stations: (1) the applicant, because of limited work experience, is unable to relate vocational and avocational interests, aptitudes, and personal goals in a meaningful way to Navy career opportunities, and (2) the recruiter is not permitted to counsel applicants into specific Navy ratings. The result is that applicants at the Recruiting Station are underinformed of their opportunities in the Navy and how the Navy can assist in realizing their career goals. Arima (1976), in his Systems Analysis of Navy Recruiting says, "The average recruiter in the field cannot possibly know enough of all jobs to counsel a prospect properly on most of them, and it is impossible for him to maintain current knowledge about the situation of openings in the near and more distant future."

The second problem occurring at the Recruiting Station is the lack of consistency of information dispensed by recruiters; even when recruiters are relating accurate information not all give the same type of information. A study by Holoter, Stehle, Conner, and Grace (1974) showed that 63% of personnel surveyed felt that their recruiters had told them only the good points of the Navy; although the information these applicants received had been accurate, it was incomplete.

The last problem is the actual accuracy of information given to applicants and the impact of inaccuracy on later reenlistment rates. Holoter et al. found that 25% of personnel surveyed felt that their recruiters had given them inaccurate information, and that this feeling was prominent among individuals who chose not to reenlist after their first term.

The problems with the recruiting interview can often lead to long term problems for the Navy. One such problem area concerns the impact the recruiting interview has on recruit training attrition. A recent study by Landau and Farkas (1978) found the following reasons for recruit training attrition: (a) nonattritees indicated that they were concerned with obtaining specific individual goals that they perceived could be attained in the Navy, (b) attritees were more vague as to goal specifications and were less likely than nonattritees to know specifically what they wanted; their decision to join the Navy occurred more from external or environmental factors than from a desire to achieve specific goals, and (c) the expectations of experiences in the Navy were more negative for attritees than for nonattritees, suggesting that attritees may not have been adequately prepared to cope with the environment of boot camp. This study suggests that current recruiting practices are a factor in the high attrition rate.

Another problem area concerns the Navy's image in the employable civilian community; individuals who become unhappy with Navy life and their jobs and choose to either attrite or leave after their first term will most likely convey a negative image of the Navy to their peers.

Scope

NPRDC is currently conducting a research project defined by the acronym Navy Personnel Accessioning System (NPAS). The primary objective of project NPAS is to develop the specifications for a computer network capable of performing four primary functions within the Navy Recruiting Command (NRC). Those functions are: (a) the optimal assignments of recruits to Navy ratings, (b) computerized adaptive testing of recruits, (c) data base management techniques and (d) computerized career guidance of recruits. The objective of this research effort is to develop and evaluate the fourth function of project NPAS, computerized career guidance of recruits. The system will be designed to support the Pre-Assignment Contacts segment of Navy Career Counseling and will provide Navy recruit applicants with uniform and unbiased information concerning Navy job opportunities and the Navy in general.

Related Research

It is generally agreed that 1908 marks the institutionalization of vocational guidance; in that year, Frank Parsons established the Vocation Bureau in Civic Service House in Boston (Ginzberg, 1971). Following that, the guidance movement generally lay dormant until the 1950's and early 60's, when it began to receive more attention due to such events as the USSR's launching of Sputnik and President Kennedy's concern with increased vocational guidance to redirect juvenile delinquents and others who were chronically unemployed.

Career guidance was seen as a means by which employment and career opportunities could be made available to a wider segment of the population. It was also viewed as a necessity for the welfare of our increasingly complex and technological society.

During the last decade the field of career guidance has profoundly felt the influence of the computer industry. Computer programs, that the designers claim are able to counsel and guide individuals in the field of education and work, have become more and more commonplace. The proponents of these programs, or more realistically systems, claim that automating certain types of career guidance functions will help solve many of the problems that have plagued the field of career guidance. Some of the specific advantages of these computer systems have been discussed by Harris and Tiedman (1974), and will not be discussed here.

Definition of Terms

The terms guidance, counseling, career, vocation, occupation, job and information have been combined in an infinite variety of ways to describe various computerized systems. This practice has caused confusion in determing exactly the functions and capabilities of each system. It is believed that if the Navy is to develop its own system it would be valuable to have it well documented and defined. The definition should begin with the systems name and how the previously mentioned terms are used in context. To avoid later confusion this research effort will adopt Super's (1976) definitions of career, vocation, occupation, and job and Ginzberg's (1971) definitions for guidance and counseling. These definitions are listed below.

Job: A group of similar, paid positions requiring some similar attributes in a single organization. Jobs are task-, outcome-, and organization-centered.

Occupation: A group of similar jobs in various organizations. Occupations are task-, economy-, and society-centered.

Vocation: An occupation with commitment, distinguished primarily by its psychological as contrasted with its economic meaning: ego-involving, meaningful to the individual as an activity, not solely for its productive, distributive, or service outcome and its economic rewards, although these too are valued. Vocations are task-, outcome-, and person-centered.

Career: The sequence of major positions occupied by a person throughout his preoccupational, occupational, and post occupational life; includes work-related roles such as those of student, employee, and pensioner, with complementary avocational, familial, and civic roles. Careers exist only as people pursue them; they are person-centered.

Guidance: Guidance includes a wide range of functions which are directed toward helping individuals make optimum use of their alternatives in acquiring an education and in pursuing a career, such as providing information and assisting in its interpretation, testing and appraising individuals, counseling, placement, and follow-up.

Counseling: Counseling, on the other hand, is a specialized function. It is usually but not always part of a guidance program. At times, counseling is the only type of assistance offered in a guidance program.

Ginzberg (1971) also provides a working definition of the term career guidance which will be adopted for the Navy system. He states that career guidance is "a process of structured intervention aimed at helping individuals to take advantage of the education, training, and occupational opportunities that are available". For the Navy system the following is also added to the above definition: Career guidance is a process by which one not only informs but also teaches the user how to explore different jobs, occupations, and vocations, what factors affect a career in terms of being successful or not, and how to make beneficial career decisions.

System Design and Development

The typical Navy applicant can be viewed as being at the stage of career development that Tiedeman (1961) refers to as exploration. A person going through this stage can be characterized as exploring a wide variety of different career options. Essentially, people are collecting information during this stage for use in later career decisions. The quantity and quality of the information that they receive during this stage will determine, to a large extent, the fortitude of their later career decisions. Thus most Navy applicants need a large amount of diverse career information concerning Navy opportunities. Unfortunately many young people are unfamiliar with the techniques for gathering this information, and, therefore, do not adequately prepare for their enlistment in the Navy.

The computerized career guidance system that is being developed for the Navy is designed to teach Navy applicants the value of career planning, help them explore different career fields, and then set short term career goals. The system will then inform them how enlistment in the Navy can help them achieve some of these career goals. An applicant would progress through the following scenario on the Navy system. First, he would be administered an adaptive screening test and an interest inventory which would be scored and interpreted for him. The applicant would then procede through a module that informs him of the value of career planning and lets him make sample career decisions. Next, the applicant would be encouraged to make some short term career goals. The system would then inform him how enlistment in the Navy could help him achieve these goals. Extensive descriptions of the ratings that the Navy would be willing to allow the applicant to enlist in would be made available to him. Including this scenario into the recruiting process, it is believed, will better prepare the applicant for a successful Navy career.

System Type

Computerized career guidance systems that have been developed in the last two decades, have been categorized into four distinct types by Harris and Tiedeman (1974). They are: (1) indirect inquiry systems--these are essentially batch-oriented system; (2) direct inquiry systems without system monitoring--these are time-sharing based systems that basically allow the user to inspect data bases with the help of instructions and code words usually available in manuals; (3) direct inquiry with system monitoring--these systems have functions that simulate a formalized type of counseling and keep track of alternatives chosen by the user; and (4) direct inquiry with system and personal monitoring--these systems are the same as the above systems but in addition they contain instruction in deliberate career decision-making and techniques for storing data about the user for use in future sessions. Preliminary research indicates that a system designed to meet the specifications of the fourth type would be of the most benefit to the Navy. Systems that have direct inquiry with system and personal monitoring are normally characterized by the following:

1. The inquirer's request for data receives almost instantaneous attention.
2. The inquirer's use of the system may be multiple or sequential, with immediate or later alteration of specified characteristics in order to produce different sets of options.
3. The inquirer is constantly aware of how his chosen characteristics are altering or diminishing his list of open options.
4. The inquirer has at his command a variety of scripts, approaches, modes, and branching opportunities which allow him a unique experience in addition to data file retrieval.
5. The data files generally can be accessed by various means and search strategies.
6. The system typically stores data about the user himself which can be meshed with data about educational and vocational options in order to generate personalized or new data.
7. The system provides some formalized counseling through procedures that monitor each session.
8. The system monitors the decision-making path of the inquirer and displays it for him and for counselors when desired.
9. The system contains instructions in deliberate career decision making.
10. The system permits the user to store data about himself to use in later career decision making.
11. The system permits the user first to personalize and later to use its original monitoring procedures to supervise his own decision making.

System Content

All computerized career guidance type systems can be divided into three basic parts: 1) the data bases - this reflects the quantity and type of information that the system can provide its users; 2) the dialogues - this reflects the techniques and ease of interacting with the system by its users; and 3) the algorithms - this indicates the type of information to be obtained from the user and how this information will be cross-referenced with the data bases. One way computerized systems can be evaluated is by comparing their characteristics on each of the three parts. The three parts of the system being developed for the Navy are discussed below.

Data Bases

Five data bases will be needed for the Navy Career Guidance System, they are: (1) Navy Ratings; (2) Civilian Occupations; (3) Navy Life Style; (4) Interest Inventory Items; and (5) Screening Test Items. A description of each is provided in the following discussion.

1. The Navy Ratings data base will include the following information: (a) title of rating; (b) general description of rating; (c) description of typical rating tasks; (d) description of working conditions; (e) aptitudes necessary for assignment to rating; and (f) relationship of rating to civilian occupations.
2. Civilian Occupations data base will include the following information on each occupation: (a) title of occupation; (b) aptitudes necessary for

entry; (c) chance of employment with a Navy background; and (d) relationship of occupation to Navy ratings.

3. The Navy Life Style data base will consist of two parts. Part 1 will consist of a general description of what one can expect to experience in the Navy from an AFEES, to recruit training, to shipboard life, to shore life, and to personal interactions with other Navy personnel. The second part will consist of a description of Navy life as experienced when assigned to each of the different ratings. This information will be provided only for the ratings that the applicant or recruit has expressed an interest in.

4. The Interest Inventory Items data base will consist of the interest inventory items to be administered to applicants. The leading candidate inventory is the Air Force's Vocational Career-Interest Examination (VOICE).

5. The Screening Test Items data base will consist of the Navy prescreening test items for administering adaptively to applicants. It is anticipated that this test will be constructed by contract.

Dialogues

The dialogues of a computerized career guidance system are the words that the user reads from the computer terminal. They are usually instructive, teaching the user how to use the system, informative, providing pertinent information to the user, or investigative, querying the user for needed information. It is anticipated that the Navy Career Guidance System will require at least the following seven dialogues: 1) the Introduction, which will orient the individual to the general career guidance system and the specific computer system, and teach him to use the computer terminal, will be precise and brief (not exceeding two or three minutes); 2) the Aptitude Test dialogue will explain the purpose of aptitude measures, teach the individual how to take the test, and administer the test, 3) the Interest Inventory dialogue will perform functions parallel to the Aptitude Test dialogue; 4) the Career Decision dialogue will instruct the individual in the value of career planning and the formulating of career decisions. The gaming part of this dialogue will permit the individual to practice making career decisions and to see how different choices he may make affect those decisions; once he has completed the gaming section, the dialogue will help him form short term career plans and goals; 5) the Navy Training dialogue will take an individual's career plans and goals as formulated by the Career Decision dialogue, correlate them with Navy ratings and civilian occupations and subsequent employability in the military and civilian communities and inform him how the Navy can help him achieve his goals. This dialogue will allow the individual to submit one or two alternative sets of plans, goals and interests to see how those changes affect the Navy's potential contributions to his anticipated career. Once this process is completed, the system will, through descriptions stored in the Navy ratings data base, permit the individual to examine the ratings appropriate for him; 6) the Navy life dialogue will utilize information from the data base of the same name to provide the individual with descriptions of Navy life as experienced by persons working in those ratings he is considering; and 7) the final dialogue of the guidance session is the Conclusion. It will summarize the information that has been provided to the individual and suggest steps he might want to take next. This dialogue will discuss: a) sources of additional information, b) the uncertainty that ratings examined at a particular guidance session will be available at a later point in time, c) that the earlier the applicant chooses to enlist the higher the probability that a rating of choice will be open, and d) a disclaimer regarding guaranteed civilian employment subsequent to a Navy career.

System Algorithms

Algorithms are the techniques employed by the system to collect information about the user, cross reference the collected information with that contained in the data oases, and develop career options for the user from the results of the cross referencing. For the Navy Career Guidance System, the following system algorithms will be required:

1. Career decision making - this algorithm will involve a gaming situation where the applicant will be encouraged to explore different career paths and opportunities by making sample career decisions. The applicant will be permitted to backtrack and change prior decisions in order to see how his overall career path changes. The goal of this algorithm will be to enable the applicant to set tentative but realistic career plans and goals.

2. Rating and occupation selection - the criteria for selecting Navy ratings and civilian occupations to be part of an applicant's career opportunities, will be the applicants aptitudes, interest measures, and career plans and goals.

3. Rating and Occupation relationship - this will be established while the data bases are being developed, and will be stored as part of the information for each rating and occupation.

4. Navy's ability to help applicant achieve career goals - this algorithm will use the data obtained from the rating and occupation selection algorithm and determine what ratings the Navy would permit the individual to enlist into, if the applicant was enlisting at the point in time that he was going through the career guidance session. The applicant will be provided information concerning how enlisting in one of the selected Navy ratings would help him achieve his career goals. The opportunity to receive detailed descriptions of each of the Navy ratings would also be provided to the applicant.

System Development

The development of a computerized career guidance system for the Navy will be accomplished through a series of stages: 1) Needs Assessment Study, 2) Data Bases, 3) System Algorithms, 4) Dialogues, 5) Delivery System, 6) Test and Evaluation, and 7) Analysis and Recommendations.

Needs Assessment Study. To document the need for a computerized career guidance system within the Navy Recruiting Command, a Needs Assessment Study of the early phases of the existing Navy career counseling system will be performed. This study will focus on:

1. The type of information currently being provided to Navy prospects, applicants, and recruits will be identified. All locations where guidance information is being provided will be included (i.e., recruiting stations, AFEES and recruit training centers). Guidance information to be reviewed will include at a minimum: (a) Navy life in general, (b) Navy careers, (c) civilian jobs and careers, (d) the Navy's potential contribution to an individual's career plans and goals, and (e) information that Navy guidance personnel should be seeking from their counselees, such as interests, hobbies, and previous work experience, as well as aptitudes and skills.

2. The specific requirements for recruiters and classifiers to present information to applicants and recruits will be documented. This will include

present Navy policy concerning actions of recruiters and classifiers, and the desire for information by the applicants and recruits. The latter will be collected by interviewing a representative sample both of recruits who have arrived at their first duty station and applicants who have chosen not to enlist in the Navy.

Data Bases. The development of the data bases will take place in four stages: 1) the number of individual data bases necessary will be determined based upon the amount and type of information to be provided by the career guidance system, 2) the size of each data base will depend upon the amount and range of information to be provided, 3) determination of the structure of the data bases (the optimal way of storing information within a data base) will be made, and 4) after completion of the previous three steps, actual construction of the data bases will proceed through the following steps: a) identification of the number and layout of specific elements of data and the interrelationships between those elements, b) data collection, c) identification of a data base management system to build and update the data bases, d) loading of the data into the data bases, and e) running test programs to insure that the data bases are properly constructed and that data can be efficiently retrieved.

System Algorithms. The system algorithms will be in part a function of the particular type of career guidance system selected. These algorithms will be based upon the state-of-the-art techniques in the field of computerized career guidance systems and will be developed after the data bases have been completed.

Dialogues. The dialogues of a computerized career guidance system are the heart of the system, since they determine how effectively the user interacts with the data bases. As such, they will be optimized in wording, length, and format and geared towards the reading level of the typical Navy applicant.

Delivery System. The system that will run the career guidance programs must have at least the following features: 1) CRT terminals, 2) printers attachable to the terminals, 3) response time under 3 seconds, 4) 10 to 15 terminals simultaneously accessing the programs, and 5) a large mass storage capability. A computer system is being procured under a NPRDC project entitled Computer-Assisted Testing, Counseling, and Assignment of Recruits, which will meet the above requirements, and which will be used to field test the career guidance programs.

Test and Evaluation. Once completed, all components of the career guidance system must be linked together and tested. The test bed site will consist of at least one Navy Recruiting District headquarters and 5 to 10 recruiting stations within that NRD. Initial testing of the system will span at least one month prior to "data collection", with changes in the programs, data bases, and dialogues being made as necessary.

The data to be collected will consist of information about system utilization, content, and career maturity measures. It will be collected by administering a questionnaire and interviews.

Analysis and Recommendations. When sufficient data has been collected, it will be compared to data collected from the Needs Assessment Study. Recommendations will include a cost/benefit analysis of the whole system and large scale implementation.

Expected Results

Benefits

The primary benefits that this research effort will have for the Navy are: a) helping to better prepare applicants for their Navy enlistment, and b) facilitating placement of the applicants into Navy job areas. These benefits will be realized because the recruits will have started to develop concrete goals for their Navy careers, and will have explored a variety of ratings that will help them achieve these goals. It is expected that by enlisting this type of recruit the Navy will experience the following benefits.

1. Recruits who will be less likely to attrite during boot camp or their first tour of duty.

2. Recruits will be less likely to request reassignment at the end of boot camp because they already know what to expect from their job.

3. Recruits will be more likely to reenlist after their first tour of duty because they will understand the value that serving in the Navy has for helping them achieve their career goals.

4. The recruits will have less of a culture shock upon arriving at their first duty station because they will know what to expect.

5. The Navy should experience better on-the-job performance from its personnel, because the recruits will not be surprised with their work requirements and will know exactly what long term benefits that they will receive from their jobs.

6. Navy personnel will more likely have rewarding experiences while in the Navy if they have a full understanding of how their Navy service will help them achieve their career goals. These personnel will also be more likely to relate their Navy experiences to their peers in a positive manner thereby enhancing the Navy's image.

7. As the Navy image becomes enhanced there is a high probability that more individuals will become interested in the Navy as an employer.

8. An enhancement in the efficiency of the Navy recruiting process due to the recruiters being relieved of having to store and transmit large amounts of information concerning Navy ratings and the Navy in general. This will free them to concentrate more on the interpersonal transactions that occur during the recruiting process. It will also allow the recruiters to spend more time giving to and collecting from the applicants pertinent information that can not be transmitted or gathered by computer.

Needed Research

It is appropriate at this point to briefly discuss some areas that are in need of additional research and that directly affect the development of a Navy computerized career guidance system. First, there does not exist a validated comparison of Navy ratings and civilian occupations. To be able to accurately suggest to a recruit civilian opportunities based upon Navy training it is imperative that such comparison be developed. The importance of this to the success of the system can be seen when, four years after full scale implementation, the first recruits, who were recruited with the automated system, leave the Navy and search for civilian employment. If they find that occupations, that they thought would be open to them, turn out not to be, due to a lack of training or inappropriate training, then they will have the right to view their Navy experience as a waste of time and the career guidance that they received as leading them astray. This situation would be no better than what currently exists and the automated system would eventually fall into disuse.

Second, it is imperative that a delivery, computer, system be available upon which to run the system. Without a delivery system that can perform all of the features of the career guidance system, the benefits will be minimized. While the development of the delivery system is the responsibility of Project NPAS, close ties will need to be kept between this research effort and Project NPAS. Any major change in the design of the delivery system could have dramatic effects upon the career guidance system.

Third, a technique will have to be developed to evaluate the effects of the Navy guidance system as a whole. There does not appear to be a standard procedure identified in the literature for this purpose. This is essentially a problem that will arise when trying to sell the system to the user, Navy Recruiting Command, but will also have an effect upon long term acceptance of the system.

Last, an organization will have to be set up to update the data bases on an annual or as needed basis. The data bases are the back bone of the guidance system. Without comprehensive, accurate, and up-to-date information the guidance system will quickly become dated and of little use. While it may at first appear that constantly updating the data bases will require an extensive maintenance staff, it is believed that streamlined procedures can be established that will keep the size of the staff to a minimum.

References

Arima, J. K. A systems analysis of Navy recruiting. Navy Postgraduate School, NPRDC SR 76-9, April 1976.

Ginzberg, E. Career Guidance. McGraw-Hill, 1971.

Harris, J. A., & Tiedman, D. V. The computer and guidance in the United past, present, and a possible future. Paper prepared for the Symposium on Computer-Based Counseling, 18th Congress, International Association of Applied Psychology, 1974.

Holoter, H. A., & Stehl, G. W., Conner, L. V., & Grace, G. L. Impact of Navy career counseling on personnel satisfaction and reenlistment Phase 2. System Development Corporation, Technical Report No. TM /003/00, 1974.

Landau, S. G., & Farkas, A. J. Selective retention: A longitudinal analysis. 1. Factors related to recruit training attrition. NPRDC TR 79-5, 1978.

Mehi, J., Holoter, H., Dow, D. S., & Grace, G. L. Preliminary description of the Navy Career Counseling Program. SDC TR NO. TM-5031/001/00, 1972.

Super, D. E. Career education and the meanings of work. U.S. Department of Health, Education, & Welfare. Stock Number 017-080-015 7, June 1976.

Tiedeman, D. V. Personnel and Guidance Journel. Sept. 1961.

PAPERS PRESENTED, LISTED BY ORDER OF AUTHOR

PAPERS PRESENTED, LISTED BY ORDER OF SUBJECT

I. PERSONNEL AND MANPOWER

1. Testing and Measurement

2. Performance Appraisal

3. Selection and Classification

4. <u>Leadership and Management Issues</u>

5. Career Information

II. PERSONNEL TRAINING

1. Curriculum Development

2. Training Evaluation

III. OCCUPATIONAL ANALYSIS

1. New Military Occupational Research Programs

2. Soft Skills / Officer Analysis

3. Development of Training Programs

4. Comprehensive Occupational Data Analysis Programs (CODAP)

STEERING COMMITTEE MEMBERS
of the
MILITARY TESTING ASSOCIATION

1. Naval Personnel Research and Development Center

2. Naval Education and Training Program Development Center

3. Army Research Institute

4. Air Force Human Resources Laboratory

5. Air Force Occupational Measurement Center

6. U. S. Coast Guard Institute

7. Canadian Forces Personnel Applied Research Unit

8. Canadian Forces Directorate for Military Occupational Structures

9. Royal Australian Air Force Evaluation Division

10. German Armed Forces Association

11. German Armed Forces Psychological Services Research Institute

REPORT OF THE

STEERING COMMITTEE MEETING

21st ANNUAL MTA

1. The Steering Committee meeting was opened with a request for nominations for the Harry H. Greer. Award. It was moved and seconded, and a unanimous vote was received to award Dr. Raymond E, Christal of the Air Force Human Resources Laboratory, San Antonio, Texas, the Harry Greer Award. 1980 nominations should be forwarded to the Chairman for the 22nd Annual Conference -

LCOL G. M. Rampton
Canadian Forces Personnel Applied Research Unit
Suite 600
4900 Yonge Street
Willowdale, Ontario
Canada M2N 6B7

2. There were no changes to the By-Laws, however there were two changes made in the membership of the Steering Committee. They were elimination of the Army Individual Training Evaluation Directorate and a change in title from Canadian Forces Directorate for Manpower Occupational Structures to Canadian Forces Directorate for Military Occupational Structures.

3. Dr. Raymond O. Waldkoetter, Army Research Institute Field Unit, Ft. Sill, Oklahoma, recommended development of a publication (in book form) which would document the major professional contributions contained in the Proceedings of previous MTA Conferences. Establishment of an Ad Hoc committee for this purpose was approved and Dr. Waldkoetter designated as Chairman.

4. LCOL G. M. Rampton and Capt. E. L. Stenton were appointed Chairman and Secretary for the next conference. The future sites for MTA conferences are:

1980	Canadian Forces Personnel Applied Research Unit (Toronto)
1981	Army Research Institute (Hampton, VA)
1982	Naval Education and Training Program Development Center (Pensacola)

HARRY H. GREER AWARD

The Military Testing Association is an outgrowth of an informal meeting of representatives of the various armed forces testing agencies in 1958. The meeting was held at the suggestion (and through the personal coordination) of CAPT Harry H. GREER, USN, Commanding Officer of the Naval Examining Center. Thus, CAPT GREER was the "founder" of the Military Testing Association. In 1962, an award in his name was created to recognize significant lasting contributions to the Association while exemplifying the ideals of the Association and its founder.

The six recipients of the award since 1962 are:

1962	CAPT Harry H. GREER, USN
1970	COL J. M. McLANATHAN, USAF
1974	MR. C. J. MacALUSO, Naval Examining Center
1977	DR. W. J. MOONAN, Naval Personnel Research and Development Center
1977	MR. J. A. BURT, U. S. Coast Guard Institute
1979	DR. Raymond E. CHRISTAL, Air Force Human Resources Laboratory

BY-LAWS OF THE MILITARY TESTING ASSOCIATION*

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel ass^ssment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

*As approved at the 1978 General Meeting of the Association 2 Nov 78, Oklahoma City, Oklahoma

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems. Each agency shall have no more than two (2) professional civilian representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

D. The Steering Committee shall have general supervicion over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. Representation from the majority of the organizations of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The Officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the association, and the Steering Committee, and to conduct official correspondence of the association, and to insure notices for conferences. The Secretary shall solicit nominations for the Harry Greer award prior to the annual conference. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment. The coordinating agencies and the order of rotation will be determined annually by the Steering Committee. The coordinating agencies for at least the following three years will be announced at the annual meeting.

C. The annual conference of the Association shall be held at a time and place determined by the coordinating agency. The membership of the association shall be informed at the annual conference of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the Chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be ex officio member of all standing committees.

B. The President with the counsel and approval of the Steering Committee may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in counsultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-third's vote of the assembled membership of the association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Enactment

These By-Laws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 2 November 1973).

1979 MTA Conferees

Norman M. Abrahams, PhD
Navy Personnel Research &
Development Center
San Diego, California 92I52

Sheila Abrams
I07II Saskatchewan Drive, #509
Edmonton, Alberta
Canada T6E 454

Homer W. Adkins
Chief of Naval Education &
Training
(CNET N-524) NAS
Pensacola, Florida 32508

Judy Akin
Commandant, US Army Infantry School
Attn: ATSH-I-V-OD (Akin)
Ft. Benning, Georgia 3I905

Robert D. Albeck
940 Aquamarine Drive
Gulf Breeze, Florida 3256I

Michele Anderson
Commandant, US Army Infantry School
ATSH-EV
Ft. Benning, Georgia 3I905

Donna C. Angle
Army Research Institute
(PTRL) KM 6532
DAR COM Bldg.
5000 Eisenhower Avenue
Alexandria, Virginia 82333

Thomas M. Ansbro
CDG CNET N-5, NAS
Pensacola, Florida 32508

Jack Anthony
Directorate of Evaluation
US Army Field Artillery School
Ft. Sill, Oklahoma 73503

Capt. Ralph H. Anzelmo, USMC
Office of Manpower Utilization
MCDEC
Quantico, Virginia 22I34

James K. Arima (54Aa)
Naval Postgraduate School
Monterey, California 93940

Paul V. Asa-Dorian
Fleet ASW Training Center, Pacific (00I)
San Diego, California 92I47

Mary A. Bachtel
Naval Education & Training
Program Development Center
Code PD-9
Pensacola, Florida 32509

Annette G. Baisden
Naval Aerospace Medical
Research Laboratory
Pensacola, Florida 32508

Harry A. Baran
AFHRL/ASR
Air Force Human Resources Laboratory
Wright-Patterson AFB, Ohio 45433

C. J. Barron
PD-I0
NET PDC
Pensacola, Florida 32509

CDR F. E. Bassett, USN
Code PD NET PDC
Pensacola, Florida 32509

ENS M. J. Bechtel
400 Oceangate, Suite 709
Long Beach, California 90822

C. Derek Beel
Officer In Charge, Naval Manpower
Utilization Unit
HMS Vernon
Portsmouth, England POI 3ER

CPT Robert R. Begland
Training Developments Institute
Attn: ATTNG-TDI-ORA
Ft. Monroe, Virginia 2365I

B. Michael Berger
DA MILPERCEN (DAPC-MSP-S)
200 Stovall Street
Alexandria, Virginia 22332

Virginia B. Berk
209 Devonshire
San Antonio, Texas 78209

JoElla Besherse
5442 Willow Cliff Road
Oklahoma City, Oklahoma 73I22

Walter W. Birdsall
Rt. 2, Box I6
Gulf Breeze, Florida 3256I

John A. Boldovici
4404 South Ridge Drive
Valley Station, Kentucky 40272

Heinz Bonn
c/o Gesellschaft fur Systementwicklung
Pferdmenges Strasse I0
D-5000 Cologne 5I
West Germany

James Boone
AAC-II8, FAA Center
P.O. Box 25082
Oklahoma City, Oklahoma 73I25

Jerome M. Booth
Hennepin County Personnel
305 Administrative Tower
Minneapolis, Minnesota 55487

Walter C. Borman
24I5 Foshay Tower
Minneapolis, Minnesota 55402

Richard D. Boyd
MILPERCEN HQDA
Alexandria, Virginia 2233I
Attn: DAPC-EPZ-HA

Edward S. Braddock
SQT Management Directorate, SIM. DIV.
US Army Training Support Center
Ft. Eustis, Virginia 23604

Laurie A. Broedling
Navy Personnel Research &
Development Center
San Diego, California 92I52

LCDR D. C. Broga
COMDT US Coast Guard
G-RT/TP54
Washington, D.C. 20590

Leland D. Brokaw, PhD
97I5 Carolwood Drive
San Antonio, Texas 782I3

Douglas J. Burrows
Education Development Division, DPCA
Bldg. 35
Ft. Benning, Georgia 3I905

John A. Burt
US Coast Guard Institute
P.O. Substation I8
Oklahoma City, Oklahoma 73I69

J. R. Byers
Data-Design Laboratories
I755 South Jeff Davis Highway, Suite 307
Arlington, Virginia 22202

Charlotte H. Campbell
Human Resources Research Organization
632 Knox Boulevard
Radcliff, Kentucky 40I60

Roy C. Campbell
Human Resources Research Organization
632 Knox Boulevard
Radcliff, Kentucky 40I60

Robert R. Carlson
28 Galt Way
Stafford, Virginia 22554

James B. Carpenter
Kentron International
Box 354I7
Brooks AFB, Texas 78235

LCDR C. W. Carter
US Coast Guard Institute
P.O. Substation I8
Oklahoma City, Oklahoma 73I69

Fred H. Casey
USAMPS/DTO (ATZN-TDP)
Ft. McClellan, Alabama 3620I

Michael J. Cassidy, SQLNDR
AFHRL/MPUS
Brooks AFB, San Antonio, Texas 78235

Raymond E. Christal
239 Killarney
San Antonio, Texas 78223

R. D. Conkwright
Westinghouse Data Score
I2310 Pinecrest Road
Reston, Virginia 22090

Ron Cooper
USA TRASANA
Bldg. I400
ATAA-TH
White Sands Missile Range
New Mexico 88002

Carl G. Cope
300 North Rawls Street
Enterprise, Alabama 36360

Bertha Cory
Army Research Institute
500I Eisenhower Avenue
Alexandria, Virginia 22333

William H. Crawford
5I95 Willow Run Drive
Pensacola, Florida 32504

John T. Dailey, PhD
80I North Pitt Street
Alexandria, Virginia 223I4

Dale M. Dannhaus, PhD
USA TRASANA
White Sands Missile Range
New Mexico 88002

Julia R. Deloney
US Coast Guard Institute
P.O. Substation I8
Oklahoma City, Oklahoma 73I69

Richard W. Dickinson
Occupational Research Program
Industrial Engineering Department
Texas A&M University
College Station, Texas 77843

Steve Dockstader, PhD
Navy Personnel Research &
Development Center
San Diego, California 92I52

Linda M. Doherty
Navy Personnel Research &
Development Center
San Diego, California 92I52

Richard Doll, PhD
3I9 North Sunset Boulevard
Gulf Breeze, Florida 3256I

Richard R Doorley
USA MILPERCEN
200 Stovall Street, Room 3507
Alexandria, Virginia 22332

Eugene H. Drucker
Human Resources Research Organization
Fort Knox Office
P.O. Box 293
Fort Knox, Kentucky 40I2I

Paul C. Duffy
Marine Corps Institute
Marine Barracks, 8th I
Box I775
Washington, D.C. 200I3

LtCol Charles V. Durham
AU/EDV
Maxwell AFB, Alabama 36II2

Joe Dwyer
Westinghouse Learning Corp.
P.O. Box 30
Iowa City, Iowa 52244

William K. Earl
ARI Field Unit (PERI-OH)
HQ TCATA
Ft. Hood, Texas 76544

Newell K. Eaton, PhD
Chief, Army Research Field Unit
Fort Knox, Kentucky 40I2I

Rolf Eckertz
c/o Gesellschaft fur Systementwicklung
Pferdmenges Strasse I0
D-5000 Cologne 50
West Germany

LTJG Gregory J. Edge
US Coast Guard Institute
P.O. Substation 18
Oklahoma City, Oklahoma 73169

John Ellis, PhD
Navy Personnel Research &
Development Center
San Diego, California 92152

Major R. T. Ellis
National Defense Headquarters
Ottawa, Ontario
Canada KIA OKZ
(Attn: DPAR-Z)

Lt. Mickey Ellison
NTTC Corry Station
Code 3331
Pensacola, Florida 32506

E. M. Evans
704 West Shorrod
Covington, Tennessee 38019

Marshall J. Farr, PhD
Director, Personnel &
Training Research
Office of Naval Research
Arlington, Virginia 22217

Eli Flyer
10 Sierra Vista Drive
Monterey, California 93940

Paul P. Foley
Navy Personnel Research &
Development Center
San Diego, California 92152

Patrick Ford
632 Knox Boulevard
Radcliff, Kentucky 40160

Robert L. Frey, Jr.
Headquarters, US Coast Guard
G-P-1/2/TP42
Washington, D.C. 20590

William A. Gager, Jr., PhD
CNET Code 015
NAS Pensacola, Florida 32508

Capt. T.J. Gallagher
Aerospace Psychology Department
NAMRL
Pensacola, Florida 32508

Jay A. Gandy
5725 MacArthur Boulevard, N.W.
Washington, D.C. 20016

Lt. R. M. Garcia
400 Oceangate, Suite 709
Long Beach, California 90822

M. J. Giorgia
AFHRL/MPS
Brooks AFB, Texas 78235

Doug Goodgame
4101 Oaklawn
Bryan, Texas 77801

Ken Gordon
952 Ocean View Avenue
Encinitas, California 92024

Steve Gorman
5102 Cliffhaven Drive
Annandale, Virginia 22003

Linda Graham
6998 Amberly Village Drive
Cordova, Tennessee 38018

LCDR John Greenfield
P.O. Box 5236
San Pedro, California 90733

Laura W. Grieger
Commandant, US Army Infantry School
Attn: ATSH-I-U-ED
Ft. Benning, Georgia 31905

Marilyn Hafer, PhD
Rehabilitation Institute
Southern Illinois University
Carbondale, Illinois 62901

Eugene R. Hall
Training Analysis & Evaluation
Group
Dept. of the Navy
Orlando, Florida 32813

Dean Halstead
Navy Personnel Research &
Development Center
San Diego, California 92152

E. Haltrecht
Personnel Research (H2-AI2)
Ontario Hydro
700 University Avenue
Toronto, Ontario
Canada M56 IX6

Randall R. Harris
HQMC (MPI-20)
Washington, D.C.

CDR F. J. Hawrysh
National Defense HQ
Ottawa, Ontario
Canada KIA OK2
Attn: DMOS 3

Robb Hayes
Oklahoma St. Merit Sytem
Jim Thorpe Memorial Bldg.
2I0I N. Lincoln Boulevard
Oklahoma City, Oklahoma 73I05

Robert G. Henderson
Commandant, Defense Language Institute
Korean Language Center
Attn: ATFL-TD-JS
Presidio of Monterey, California 93940

Richard E. Hiiligoss
Army Research Institute
P.O. Box 2086
Bldg. 75D
Ft. Benning, Georgia 3I905

Lt. James A. Hodgdon
3266 Ashford Street, Apt. D
San Diego, California 92III

CDR John D. Holland
NODAC, Bldg. I50
Washington Navy Yard (Anacostia)
Washington, D.C. 20374

Charles W. Howard, PhD
US/ARI
P.O. Box 6057
Ft. Bliss, Texas 799I6

Judith Huffman
Fleet ASW Training Center, PAC
Code 262
San Diego, California 92I47

LCDR I. L. Jackson
National Defense HQ
Ottawa, Ontario
Canada KIA OK2
Attn: DMOS 3-2

J.B. Joaquin
Ontario Hydro
700 University Avenue
Toronto, Ontario
Canada M5G IX6

Robert N. Johnson
US Army ADMINCEN
Ft. Ben, Harrison
Indianapolis, Indiana 462I6

Karen N. Jones
US Coast Guard Institute
P.O. Substation I8
Oklahoma City, Oklahoma 73I69

Carolyn H. Josey
Lake Taylor High School
I384 Kempsville Road
Norfolk, Virginia 23502

Major Grover A. Josey, Jr.
II44 Janaf Place
Norfolk, Virginia 23502

Jeffrey Kantor, PhD
AFHRL/MPUF
Brooks AFB, Texas 78236

Bert King
Code 452
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 222I7

Constance L. Kintop
6II4 Elliot Avenue South
Minneapolis, Minnesota 554I7

Peggy Knaup
US Coast Guard Institute
Deck Branch
P.O. Substation I8
Oklahoma City, Oklahoma 73I69

C. Mazie Knerr, PhD
Litton Mellonics
P.O. Box I286
Springfield, Virginia 22I5I

Major A. N. Knox
I26 Belle Drive
Marina, California 93933

Major Kobes
I2 Seaborn
Midland City, Alabama 36350

Arthur L. Korotkin, PhD
Institute for Behavioral Research
2429 Linden Lane
Silver Spring, Maryland 209I0

CPT Dennis M. Kowal, PhD
Walter Reed Army Medical Center
Psychiatry Service
Washington, D.C. 200I2

Burton Krain, PhD
US Office Personnel Management
230 South Dearborn
Chicago, Illinois 60604

Leonard Kroeker, PhD
Navy Personnel Research &
Development Center
San Diego, California 92I52

Major C. D. Kuhn
HQMC (MPI-20)
Washington, D.C. 20380

Robert E. Lambert
Professional Associate
Educational Testing Services
1947 Center Street
Berkeley, California 94704

Richard S. Lanterman
HQ US Coast Guard
G-P-I/2/TP42
Washington, D.C. 20590

Virginia L. Lee
Army Education Center
Ft. Ord, California 9394I

LtCol C. A. Leech
Director Military Occupational
Structures
National Defense Headquarters
Ottawa, Ontario
Canada KIA OK2

Nancy A. Lewis
AFHRL/TSZ
Brooks AFB, Texas 78235

LCDR Robert J. Low
DNTR (Navy Office)
c/o Australian Embassy
I60I Massachusetts Avenue
Washington, D.C.

Milton Maier
Army Research Institute
500I Eisenhower Avenue
Alexandrai, Virginia 22333

Owen Maller, PhD
Behavioral Sciences Division
US Army R&D CMMD
Matick, Maryland 0I760

Ruth Ann Marco
McDonnell Douglas Astronautics Co.
Engineering Psychology Department
P.O. Box 5I6
St. Louis, Missouri 63I66

J. I. Markowitz
Commandant, US Coast Guard
Attn: BAE-3
Washington, D.C. 20590

John F. McAreavy, PhD
I03I Newell Avenue
Muscatine, Iowa 5276I

R. E. McCutcheon, Jr.
FLECOMBATRACENPAC
Code OOE
200 Catalina Boulevard
San Diego, California 92I47

LtCol W. W. McIver, USMC
Assistant Director
Office of Manpower Utilization, MCDEC
Quantico, Virginia 22I34

John T. Meehan
AULEDV
Maxwell AFIS, Alabama 36II2

J. B. Meredith, Jr., PhD
P.O. Box I2773
Norfolk, Virginia 23502

Donald J. Metz
National Computer Systems, Inc.
P.O. Box 6875
476 Georgetown Avenue
Ventura, California 93003

Nancy Mitchell, PhD
Commandant, US Army Infantry School
ATSH-EV
Ft. Benning, Georgia 3I905

Amelia E. Mobley
US Coast Guard (G-PMR-5/TP44)
2I00 2nd Street S.W.
Washington, D.C.

John B. Mocharnuk, PhD
McDonnell Douglas Astronautics Co.
Engineering Psychology Department
P.O. Box 5I6
St. Louis, Missouri 63I66

Thomas Molloy
4603 Vance Jackson, #I404
San Antonio, Texas 78230

Brian E. Moore
College of Business
University of Texas, Austin
Austin, Texas 78704

B. E. Moyer
FLTCOMBATRACENPAC
Code 00E2
200 Catalina Boulevard
San Diego, California 92I47

Stephen J. Mussio
Minneapolis Personnel Department
3I2 3rd Avenue South
Minneapolis, Minnesota 554I5

Daniel J. Naert
Rock Island Arsenal
Training & Development Division
SARRI-PTT
Rock Island, Illinois 6I299

Ray W. Nelson
7439 Wymalt Road
Pensacola, Florida 32506

Nancy Nieboer, PhD
HQ USAREC
Attn: USARCASP-E
Ft. Sheridan, Illinois 60037

William L. Osborn
Human Resources Research Organization
P.O. Box 293
Fort Knox, Kentucky 40I2I

James E. Osborne
PD3
USNETPC
Pensacola, Florida 32509

Ronald C. Page, PhD
Corporate Compensation
Control Data Corporation
HQN06V
8I00 34th Avenue South
Minneapolis, Minnesota 55440

Linda Pappas
9804 Ward Court
Fairfax, Virginia 22030

Lorraine Penfold
COMDT (G-RT), US Coast Guard
Washington, D.C. 20003

E. N. Pickrel, PhD
AAM-520
Office of Aviation Medicine
Federal Aviation Administration
800 Independence S.W.
Washington, D.C. 2059I

C. M. Pipkin
4305 Cheltenham Circle
Pensacola, Florida 32504

Major Peter Plaut
The Judge Advocate General's School
Charlottesville, Virginia 2290I

LCDR Earl H. Potter III
Department of Humanities
US Coast Guard Academy
New London, Connecticut 06320

Klaus J. Puzicha, PhD
Dezernat Wehrpsychologie
im Streitkrafteamt
Postfach 20 50 03
D-5300 Bonn 2
West Germany

Col. R. J. Rabin
USA TRASANA
White Sands Missile Range
New Mexico 88002

Elizabeth M. Ralls, PhD
USA TRASANA
Attn: ATAA-TH
White Sands Missile Range
New Mexico 88002

LtCol G. M. Rampton, PhD
4900 Yonge Street, Suite 600
Willowdale, Ontario
Canada M2N 6B7

George D. Rastall
CNET, Code NIOI
NAS
Pensacola, Florida 32508

John H. Rathkamp
Army Education Center
Bldg. II2
Ft. McPherson, Georgia 30330

Malcolm J. Ree, PhD
AFHRL/MP
Brooks AFB, Texas 78249

Paul A. Reed
8I8 S. Park Drive
Petersburg, Virginia 23803

LCDR William Reinhardt
NODAC
Bldg. I50, WNY, Anacostia
Washington, D.C. 20374

Donald B. Rock, APT-I
Federal Aviation Administration
800 Independence Avenue
Washington, D.C.

Marty Rockway, PhD
Technical Director
AFHRL/TT
Lowry AFB, Colorado 800I2

FKpt. R.-E. Rolfs
BMVG-FUE SI3
Postfach I328
5300 Bonn I
West Germany

Capt. P. Rossiter
National Defense HQ
Ottawa, Ontario
Canada KIA OK2
Attn: DMOS 3-4

Hendrick W. Ruck
AFHRL/MPUS
Brooks AFB, Texas 78235

Gregory Runyan
I90I Okaloosa Street
Avalon Beach, Florida 32570

J. R. Rush
NETPDC, Saufley
Pensacola, Florida 32509

Sydney Sako
OTS/MTCM
Lackland AFB, Texas 78236

William A. Sands
Navy Personnel Research &
Development Center
San Diego, California 92I52

CDR Greg J. Sanok
US Coast Guard Training Center
Alameda, California 9450I

Dorothy Scanland, PhD
DANTES
Pensacola, Florida 23509

Worth Scanland
CNET (N-5)
Pensacola, Florida 32508

Rich Schram
NTTC Corry Station
Code I0232
Pensacola, Florida 325II

Leonard C. Seeley
450I Sellman Road
Beltsville, Maryland 20705

S. B. Sells, PhD
Institute of Behavioral Research
Texas Christian University
Fort Worth, Texas 76I29

David E. Servinsky
7872 North Road
Severn, Maryland 2II44

H. E. Seuberlich
Sued Strasse I23
D-53 Bonn 2
West Germany

Victor Shaw
P.O. Box 5822
Presidio of Monterey, California 93940

CPT David L. Sheets
3723 Chartwell Drive
San Antonio, Texas 78230

Albert E. Shively
4340 Burtonwood Road
Pensacola, Florida 32504

William R. Shoen
SR. ED. Advisor
Code 02A
Service School Command
NAVATTRNGCNT
Orlando, Florida 328I3

Edgar L. Shriver, PhD
I500 N. Beauregard Street, Suite 205
Alexandria, Virginia 223II

William H. Sims
Center for Naval Analyses
2000 N. Beauregard Street
Alexandria, Virginia 223II

John P. Smith, PhD
I383I Durango Drive
Del Mar, California 920I4

J. W. Smith
Commandant, US Army Missile &
Munitions Center & School
Attn: ATSK-TD-AD-L
Redstone Arsenal, Alabama 35809

Margaret J. Smith
3I6 Interbay Avenue
Pensacola, Florida 32507

E. Solomon
8460 Old Spanish Trail
Pensacola, Florida 32504

Daniel E. Spector
I6I5 Fairway Circle
Jacksonville, Alabama 36265

Jimmie B. Spivey
50I Slaters Lane, #I005
Alexandria, Virginia 223I4

Capt. E. L. Stenton
4900 Yonge Street, Suite 600
Willowdale, Ontario
Canada M2N 6B7

Robert W. Stephenson, PhD
AFHRL/AE
Brooks AFB, Texas 78235

Major Stanley D. Stephenson
AFMPC/MPCYPT
Randolph AFB, Texas 78I48

Dale W. Stewart
Rt. 4, Box I80
Elizabethtown, Kentucky 4270I

Michael C. Thew
AFHRL/TSPZ
Brooks AFB, Texas 78235

Burt E. Thompson
435I Burtonwood Drive
Pensacola, Florida 32504

Nancy Thompson
AFHRL/MPUS
Brooks AFB, Texas 78235

John D. Tubbs
USA TRASANA, ATAA-THC
White Sands Missile Range
New Mexico 88002

Lt. Pamela W. Tubbs
3832 Belleau Drive
Memphis, Tennessee 38I27

Joseph A. Tucker, PhD
Air Traffic Controllers
I425 Fern Oak Court
McLean, Virginia 22I0I

Thomas C. Tuttle, PhD
Director, Maryland Center for
Productivity & Quality of Working Life
University of Maryland
College Park, Maryland 20742

J. E. Uhlaner, VP, PhD
Perceptronics, Inc.
627I Variel Avenue
Woodland Hills, California 9I367

Sally J. Van Nostrand
US Army Research Institute
500I Eisenhower Avenue
Alexandria, Virginia 22333

Phyllis Voorhees
Coast Guard Institute
P.O. Substation I8
Oklahoma City, Oklahoma 73I69

Raymond O. Waldkoetter
ARI Field Unit
P.O. Box 33066
Ft. Sill, Oklahoma 73503

LTC Bradford L. Walton
HQ TRADOC
ATTNG-TDI-ORAD
Ft. Monroe, Virginia 2365I

Joe H. Ward, Jr., PhD
AFHRL
Brooks AFB, Texas 78235

Thomas A. Warm
US Coast Guard Institute
P.O. Substation I8
Oklahoma City, Oklahoma 73I69

LtCol Brian K. Waters
AWC/EDV
Maxwell AFB, Alabama 36II2

Tom Watson
Air Force Human Resource Laboratories
AFHRL/MP
Brooks AFB
San Antonio, Texas 78235

Johnny J. Weissmuller
AFHRL/TSPZ
Brooks AFB, Texas 78235

Dr. Welch
Commander, US Army Training Support
Center
Attn: ATTSC-AI-PO
Ft. Eustis, Virginia 23604

Capt. John R. Welsh
Air Force Manpower &
Personnel Center
Randolph AFB, Texas 78I48

James D. Wiggins
406 Shoreline Drive
Gulf Breeze, Florida 3256I

Capt. D. Mark Wilkinson
AIFOS/EOV
Bldg. 803
Maxwell AFB, Alabama 36II2

Donald H. Williams
National Computer Systems, Inc.
440I West 76th Street
Minneapolis, Minnesota 55435

Rayburn A. Williams
Code N53
Chief of Naval Education & Training
Pensacola, Florida 32508

Richard Willing
FAA Center
P.O. Substation I8
Oklahoma City, Oklahoma 73I69

Major Robert Wiltshire
CG USA SIG CEN & Ft. Gordon
Ft. Gordon, Georgia 30909

Martin Wiskoff, PhD
Navy Personnel Research &
Development Center
San Diego, California 92I52

Darrell A. Worstine
Commander, US Army MILPERCEN
Attn: DAPC-MSP-D
200 Stovall Street
Alexandria, Virginia 2233I

W. H. Wulfeck, PhD
Code P304
Navy Personnel Research &
  Development Center
San Diego, California 92I52

Charles Yackulic
Seattle University
Seattle, Washington 98I22

Ted Yellen
Navy Personnel Research &
  Devlopment Center
San Diego, California 92I52

Robert W. Youtt
Army Education Center
T2207
Ft. Devons, Massuchusetts 0I433

Henry R. Ziel
82I0 - III Street, #III2
Edmonton, Alberta
Canada T6G 207

Major R. A. Zuliani
CFPARU
4900 Yonge Street
Willowdale, Ontario
Canada M2N 6B7